Award Number: W81-XWH-09-2-0175

TITLE: Development of EBM-CDSS (Evidence-Based Clinical Decision Support System) to AIG Prognostication in Terminally Ill Patients".

PRINCIPAL INVESTIGATOR: Benjamin Djulbegovic, MD, PhD

CONTRACTING ORGANIZATION: University of South Florida
Tampa, FL 33612

REPORT DATE: March 2016

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188
1. REPORT DATE *(DD-MM-YYYY)* March 2016	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* 25Sep2009 - 31Dec2015
4. TITLE AND SUBTITLE Development of EBM-CDSS (Evidence-Based Clinical Decision Support System) to AIG Prognostication in Terminally Ill Patients".		**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** W81-XWH-09-2-0175 **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Dr. Djulbegovic Benjamin email: bdjulbeg@health.usf.edu		**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of South Florida 4202 E. Fowler Avenue, Tampa, FL 33620		**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012		**10. SPONSOR/MONITOR'S ACRONYM(S)** **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution unlimited		
13. SUPPLEMENTARY NOTES		

14. ABSTRACT

Goal of the project is to develop an Evidence-based Clinical Decision Support (CDSS-EBM) system and make it available at the point of care to improve prognostication of the life expectancy of terminally ill patients to improve referral of patients to hospice. In addition, the EBM-CDSS will be expanded with an evidence based pain management module (EB-PMM) to assist physicians managing patients with pain.

The study was conducted at the Moffitt Cancer Center (MCC) and Tampa General Hospital (TGH). [Both sites received scientific review committee, IRB and the sponsor approval]. The study was successfully closed on December 31, 2015. Our key research accomplishments are as follows:

- We have successfully designed Evidence-based Clinical Decision Support System (CDSS-EBM) software to facilitate end of life care decisions. It will be available on the web once the final publication is available in the public domain (the paper is submitted for publication).
- We enrolled 184 study participants [we have screened and approached 1052 patients]
- We used both the English and Spanish version of the informed consent forms for our study.
- The final CDSS-EBM for hospice referral will be made available along with evidence-based pain management module, which we also developed during as a part of this project.
- We have published numerous manuscripts in peer-reviewed journals as well as abstracts at prestigious national meetings. Our key findings were presented at the Annual Meeting of Society of Medical Decision-Making held in St Louise in October, 2015. The paper was rated as the second-best submission (the final report is submitted for publication).

15. SUBJECT TERMS

CDSS, SUPPORT, DEALE, Terminally ill, Hospice, Prognostication

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	223	19b. TELEPHONE NUMBER *(include area code)*

Table of contents	
Item description	Page number

Introduction

The goal of this project was to develop an Evidence-based Clinical Decision Support System (CDSS-EBM) available at the point of care which will improve prognostication of life expectancy of terminally ill patients and facilitate the hospice referral process. In addition, the CDSS-EBM was expanded with an evidence based pain management module (EB-PMM) to assist physicians managing patients with pain.

Body:

Key research-related accomplishments:

- We screened 1052 participants for eligibility and finally enrolled a total of 184 study participants.
- The study was available to both English and Spanish speaking patients.
- We have successfully designed CDSS-EBM software to facilitate end of life care decisions. Features of the software include: utilization of multiple prognostication models, incorporation of the dual visual analogue scales for elicitation of regret, elicitation of acceptable regret, incorporation of treatment effects in the decision making calculations. The details of the CDSS-EBM are published in a peer-reviewed journal manuscript (See appendix: Extensions to Regret-based Decision Curve Analysis: An application to hospice referral for terminal patients. BMC medical informatics and decision making, 11(1), 1, 2011). The final version will be made available in the public domain once the paper, which is submitted for publication, is accepted for publication.
- We have also successfully designed Evidence-based Chronic Pain Management Module (EB-PMM) to complement the CDSS-EBM. The version is available in JAVA language (web version) and for IpAD. The final version will be made available in the public domain once the paper, which is submitted for publication, is accepted for publication.
- Over the period of the project we obtained approvals in terms of continuing review reports from the University of South Florida (USF) Institutional Review Board on regular basis and never missed a deadline. The study will be closed at IRB after the final report is published.

Reportable outcomes

Publications in peer reviewed journals: (attached in appendices)

Hozo I, Djulbegovic B, Luan S, Tsalatsanis A, Gigerenzer G. Towards theory integration: Threshold model as a link between signal detection theory, fast-and-frugal trees and evidence accumulation theory. Journal of evaluation in clinical practice. 2016 Jan 1.

Cucchetti A, Djulbegovic B, Tsalatsanis A, Vitale A, Hozo I, Piscaglia F, Cescon M, Ercolani G, Tuci F, Cillo U, Pinna AD. When to perform hepatic resection for intermediate-stage hepatocellular carcinoma. Hepatology. 2015 Mar 1;61(3):905-14.

Gil-Herrera E, Aden-Buie G, Yalcin A, Tsalatsanis A, Barnes LE, Djulbegovic B. Rough set theory based prognostic classification models for hospice referral. BMC medical informatics and decision making. 2015 Nov 25;15(1):98.

Djulbegovic B, Tsalatsanis A, Hozo I. Determining optimal threshold for statins prescribing: individualization of statins treatment for primary prevention of cardiovascular disease. Journal of evaluation in clinical practice. 2015 Dec 1.

Djulbegovic B, Elqayam S, Reljic T, Hozo I, Miladinovic B, Tsalatsanis A, Kumar A, Beckstead J, Taylor S, Cannon-Bowers J. How do physicians decide to treat: an empirical evaluation of the threshold model. BMC medical informatics and decision making. 2014 Jun 5;14(1):1.

Hernandez JM, Tsalatsanis A, Humphries LA, Miladinovic B, Djulbegovic B, Velanovich V. Defining optimum treatment of patients with pancreatic adenocarcinoma using regret-based decision curve analysis. Annals of surgery. 2014 Jun 1;259(6):1208-14.

Djulbegovic B, Beckstead JW, Elqayam S, Reljic T, Hozo I, Kumar A, Cannon-Bowers J, Taylor S, Tsalatsanis A, Turner B, Paidas C. Evaluation of physicians' cognitive styles. Medical Decision Making. 2014 Jul 1;34(5):627-37.

Gil-Herrera E, Tsalatsanis A, Kumar A, Mhaskar R, Miladinovic B, Yalcin A, Djulbegovic B. Identifying homogenous subgroups for individual patient meta-analysis based on Rough Set Theory. In Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE 2014 Aug 26 (pp. 3434-3437). IEEE.

Wao, H. O., Mhaskar, S. R., Kumar, A., Miladinovic, B., Guterbock, T., Hozo, I., & Djulbegovic, B. (2014). Uncertainty about effects is a key factor influencing Institutional Review Boards' approval of clinical studies. Annals of Epidemiology, 24(10), 734-740.

Miladinovic B, Mhaskar R, Kumar A, Kim S, Schonwetter R, Djulbegovic B, “External Validation of a Web-based Prognostic Tool for Predicting Survival for Patients in Hospice care”, J Palliat Care. 2013, 29(3):140-6. PMID: 24380212.

Wao H., Mhaskar, R., Kumar, A., Miladinovic, B., & Djulbegovic, B. (2013). Survival of patients with non-small cell lung cancer without treatment: a systematic review and meta-analysis. Systematic Reviews, 4:2(1), 115-134.

Miladinovic B, Kumar A, Mhaskar R, Kim S, Schonwetter R, Djulbegovic B, “A flexible alternative to the Cox proportional hazards model for assessing the prognostic accuracy of hospice patient survival”, PLoS One. 2012, 7(10):e47804. PMID: 23082220; PMCID: PMC3474724.

Gil-Herrera E, Yalcin A, Tsalatsanis A, Barnes LE, Djulbegovic B. Towards a classification model to identify hospice candidates in terminally ill patients. In Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE 2012 Aug 28 (pp. 1278-1281). IEEE.

Djulbegovic B, Hozo I, Beckstead J, Tsalatsanis A, Pauker SG. Dual processing model of medical decision-making. BMC medical informatics and Decision Making. 2012 Sep 3;12(1):94.

Tsalatsanis A, Barnes LE, Hozo I, Djulbegovic B. Extensions to Regret-based Decision Curve Analysis: An application to hospice referral for terminal patients. BMC medical informatics and decision making. 2011 Dec 23;11(1):1.

Gil-Herrera E, Yalcin A, Tsalatsanis A, Barnes LE, Djulbegovic B. Rough Set Theory based prognostication of life expectancy for terminally ill patients. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE 2011 Aug 30 (pp. 6438-6441). IEEE.

Tsalatsanis A, Hozo I, Vickers A, Djulbegovic B. A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making. BMC medical informatics and decision making. 2010 Sep 16;10(1):51.

Meeting abstracts: (attached in appendices)

Tsalatsanis A, Hozo I, Djulbegovic B. Empirical testing of regret-based threshold model in the end-of-life care. 37th Annual Meeting of Society of Medical Decision-Making, October 18-21, 2015 St Louis, MO.

Tsalatsanis A, Hozo I, Djulbegovic B. Empirical evaluation of regret and acceptable regret model. 36th Annual Meeting of Society of Medical Decision Making, Miami, October 19-22, 2014.

J.M. Hernandez, A. Tsalatsanis, B. Djulbegovic, V. Velanovich, “Regret theory modeling in pancreatic adenocarcinoma” (poster). Annual Cancer Symposium of the Society of Surgical Oncology Orlando, Florida, Mar 21-24, 2012.

R. Mhaskar, B. Miladinovic, A. Tsalatsanis, A. Mbah, A. Kumar, K. Sehwan, R. Schonwetter, B. Djulbegovic, “External validation of prognostic models in terminally ill patients” (poster). ASH Annual Meeting and Expo, San Diego, California, Dec 10-13, 2011. Abstract published in Blood, vol. 118 (21), 2011.

I. Hozo, A. Tsalatsanis, A. Vickers, B. Djulbegovic, “A regret theory approach to decision curve analysis” (poster). Annual Meeting of Society for Medical Decision Making (SMDM), Toronto, Canada, Oct 18-21, 2010.

B. Djulbegovic, J. Beckstead, A. Tsalatsanis, R. Mhaskar, A. Flynn, O. Fabelo, H. Tuch, A. Kumar, E. Pathak, I. Hozo, P. Jacobsen, “Anticipatory regret of commission but not omission leads to low post-decisional regret in terminally ill patients” (oral). Biennial European Meeting of SMDM.

Conclusions

Main conclusions of our work include:

1. We have developed a novel method for eliciting decision maker's preferences based on regret theory, and which can facilitate decision-making in the end-of-life setting. We showed that the method accurately represent true patients' preferences 85% of time and predict actual choice (continue treatment vs. choosing hospice) 71% of times. We have developed new tool (decision support system) , which can be easily integrated within clinical work-flow.
2. We have applied our method successfully across all disease and have also have modified it and further tested it in the management of specific diseases including such intermediate HCC, pancreatic cancer, and statin use.
3. We have extended our regret approach to derive the first dual processing model of medical decision-making that has potential to enrich the current medical decision-making field.
4. We have validated survival models using a novel family of flexible survival functions in the context of a web-based prediction tool.
5. We showed that a flexible family of survival models predicts survival more accurately that the commonly used Cox proportional hazards model, by allowing for the flexible modeling of the baseline survival function.
6. We performed an external validation of a web-based interactive prognostic tool using novel survival models. We established that for a model to be useful to hospice and palliative care researchers, it should report explicit risk scores and estimates of baseline survival to be combined with new patient information to provide guidance on how this should be done.
7. Finally, and most importantly, we have successfully explored our theoretical framework to successfully develop new tools (software application) to facilitate decision-making in the end-of-life setting and hospice referral process.
8. We have also developed a new theoretical framework about which we are particularly excited- we linked our regret threshold model with signal detection theory, fast-and-frugal heuristics and evidence accumulation theory- and which we believe hold promise for decision-making applications across diverse medical settings.

Three manuscripts outlining our overall study findings are being submitted them for publication in peer-reviewed journals.

The titles of these manuscripts are as below:

- Eliciting people's regret improves the decision-making at the end of life
- Active treatment in the end-of-life setting does not prolong survival in comparison to palliative care and is associated with increased toxicity. A systematic review
- External validation of performance of palliative performance scale (PPS) and modified PPS

Next steps:

Once all our publications are in the public domains, we hope to apply for funding from state and national agencies including the Department of Defense to continue our research in this field including scaling up our findings across diverse settings and institutions.

Appendix

- **Includes PDFs of all publications to date**

(provided in separate files)

Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

Towards theory integration: Threshold model as a link between signal detection theory, fast-and-frugal trees and evidence accumulation theory

Iztok Hozo PhD,[1] Benjamin Djulbegovic MD PhD,[3,4,5] Shenghua Luan PhD,[6] Athanasios Tsalatsanis PhD[2] and Gerd Gigerenzer PhD[5]

[1]Professor, Department of Mathematics, Indiana University, Gary, IN, USA
[2]Associate Professor, USF Health Program for Comparative Effectiveness Research, Division for Evidence-Based Medicine, Department of Internal Medicine, University of South Florida, Tampa, FL, USA
[3]Professor of Oncology, Departments of Hematology and Health Outcome Behavior, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA
[4]Director of Medical Research, Tampa General Hospital, Tampa, FL, USA
[5]Professor, [6]Researcher, Max Planck Institute for Human Development, Berlin, Germany

Keywords
clinical guidelines, evaluation, evidence-based medicine

Correspondence
Professor Benjamin Djulbegovic
12901 Bruce B. Downs Blvd, MDC02
Tampa, FL 33612
USA
E-mail: bdjulbeg@health.usf.edu

Accepted for publication: 6 November 2015

doi:10.1111/jep.12490

Abstract

Rationale, aims and objectives Theories of decision making are divided between those aiming to help decision makers in the real, 'large' world and those who study decisions in idealized 'small' world settings. For the most part, these large- and small-world decision theories remain disconnected.

Methods We linked the small-world decision theoretic concepts of signal detection theory (SDT) and evidence accumulation theory (EAT) to the threshold model and the large world of heuristic decision making that rely on fast-and-frugal decision trees (FFT).

Results We connected these large- and small-world theories by demonstrating that seemingly different decision-making concepts are actually equivalent. In doing so, we were able (1) to link the threshold model to EAT and FFT, thereby creating decision criteria that take into account both the classification accuracy of FFT and the consequences built in the threshold model; (2) to demonstrate how threshold criteria can be used as a strategy for optimal selection of cues when constructing FFT; and (3) to show that the compensatory strategy expressed in the threshold model can be linked to a non-compensatory FFT approach to decision making. We also showed how construction and performance of FFT depend on having reliable information – the results were highly sensitive to the estimates of benefits and harms of health interventions. We illustrate the practical usefulness of our analysis by describing an FFT we developed for prescribing statins for primary prevention of cardiovascular disease.

Conclusions By linking SDT and EAT to the compensatory threshold model and to non-compensatory heuristic decision making (FFT), we showed how these two decision strategies are ultimately linked within a broader theoretical framework and thereby respond to calls for integrating decision theory paradigms.

Introduction

Theories of decision making deal with either 'large-' or 'small'-world phenomena [1]. In a small world, decision makers are not under time pressure and have access to all relevant knowledge, including all alternatives, consequences and probabilities. In turn, such knowledge enables a decision maker to make an optimal, rational decision. A prototype of 'small'-world theory is expected utility theory (EUT). By contrast, in a 'large' (i.e. typically a real) world, knowledge about the complete set of alternatives, consequences and probabilities is limited. Under these circumstances, a rational decision maker relies on adaptive cognitive processes that surprisingly often lead to efficient and accurate choices. A prototype of large-world theory is the heuristic theory of decision

making [2]. Until lately, no attempt had been made to connect decision theories of large worlds with those of small ones. Recently, however, Luan *et al.* successfully applied small-world signal detection theory (SDT) to fast-and-frugal tree (FFT) heuristics [3]. In a separate undertaking, Lee and Cummins [4] proposed evidence accumulation theory (EAT) as a concept unifying heuristics and 'rational' models.

In a further endeavor to unify different psychological and statistical theories [5], we link the threshold model (formulated within both the EUT and regret framework) [6–8] with SDT and FFT and with EAT. We illustrate the usefulness of this new framework in the context of medical decision making with the example of prescribing statins for primary prevention of cardiovascular disease (CVD).

Methods

SDT

SDT, which has its origins in the Neyman–Pearson theory of hypothesis testing, is widely used throughout science and medicine [3]. Its fundamental assumption resides on the notion that the two possible events (*signal*, e.g. presence of disease, and *noise*, e.g. absence of disease) have overlapping distributions on an observation scale X [9]. Each of these distributions is further divided into two possible outcomes, which are determined by setting a decision criterion (x_c, as in Fig. 1). The criterion divides the signal distribution into true positives (hits) and false negatives (misses). In medicine, the hit rate is called *sensitivity* (S) and

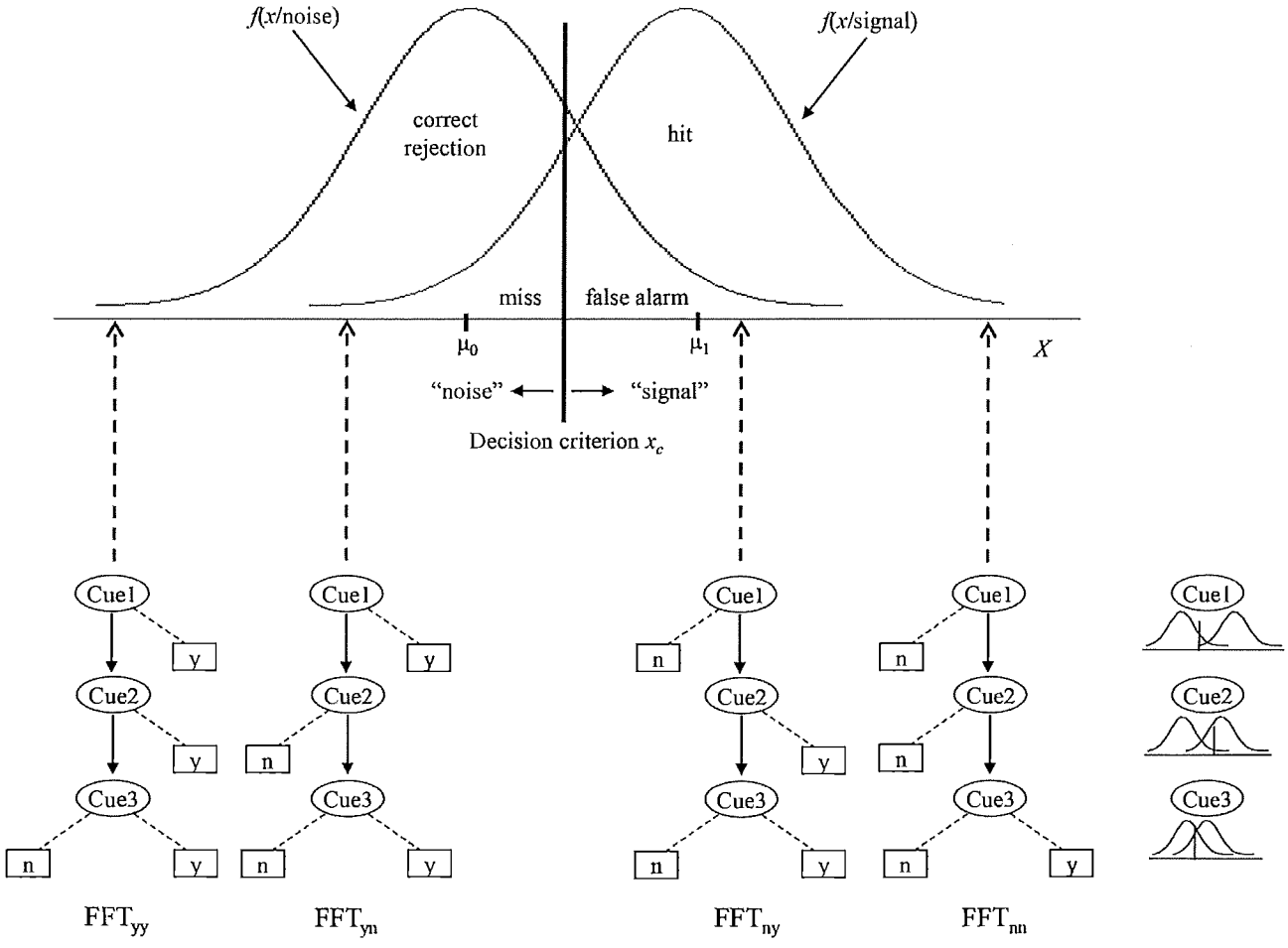


Figure 1 A relationship between signal detection theory (SDT) and fast-and-frugal trees (FFT). The upper part of the figure illustrates the concepts of SDT in a binary decision task, and the lower part illustrates the four possible FFTs that can be constructed when three cues are searched in a set order. Based on the decisions pointed to by the first two exits, the trees are named from left to right FFTyy, FFTyn, FFTny and FFTnn (where y stands for 'yes' and n for 'no'). The arrows connecting the figure parts indicate the rough locations of the four FFTs' decision criteria when they are used to make a binary y/n (for signal and noise, respectively) decision. Among the four, FFTyy has the most liberal decision criterion, and FFTnn the most conservative one. The decision criteria of FFTyn and FFTny are less extreme than the other two, with FFTyn being more liberal than FFTny. The two overlapping normal distributions next to each cue illustrate SDT's assumption of how object values are distributed on a cue and emphasize that each cue comes with its own discriminability and decision criterion (in *d'* and *c*, respectively; see Table 1 – Appendix). Note that the structure of FFTs shown in the figure is not based on contrasting it to the threshold (the latter is shown in Fig. 5; see text for further explanations; figure is based on Luan *et al.* [3]).

1 – sensitivity is the *false-negative rate* (FN). The noise distribution is composed of true negatives (correct rejections) and false positives; their rates are called *specificity* (C) and *false alarm rate* (FA), respectively [9]. From these data, we can calculate the accuracy of classification of a particular diagnostic cue. Because the consequences of misses versus false alarms often differ, two additional metrics are typically calculated according to SDT. These are *d'* (discrimination) and *c* (decision criterion) [9]. Discrimination or discriminability (*d'*) measures the distance between the means of signal and noise in standard deviation units (*z*-scores). The decision criterion (*c*) (also known as *response bias* [3,9]) determines the decision cut-off; if it is set at $c = 0$, then *FA* and *FN* are weighted according to the prior probabilities of signal and noise. If *c* is moved to the left (<0), then $FN < FA$ relative to prior probabilities of signal and noise (*liberal bias*). The opposite holds when *c* is moved to the right (>0). In this case, *FN*s are more tolerated than *FA*s ($FN > FA$), relative to prior probabilities of signal and noise (*conservative bias*). The threshold likelihood ratio of the decision at hand according to SDT is given as:

$$\beta_{optimal} = \frac{V(C) - V(FA)}{V(S) - V(FN)} \cdot \frac{p(Noise)}{p(Signal)} \quad (1)$$

where $V(\cdot)$ represents the utility of a particular outcome [3] (see section on threshold model below). For terminology of SDT, see Appendix A.

FFTs

FFTs are a class of simple heuristic decision-making strategies that relies on limited information to reduce estimation error and facilitate fast decisions [1,2,10]. An FFT is a decision tree composed of sequentially ordered cues. Typically, cues and decisions are binary (yes/no), and their relation can be framed as *if–then* statements (e.g. if a person has severe chest pain, then perform diagnostic tests to rule out myocardial infarction). If the condition is met, the decision can be made and the FFT is exited. If the condition is not met, the FFT considers the other cues, one after another, until the exit condition of a cue is met. The last cue of an FFT has two exits to ensure that a decision is ultimately made [3]. Formally, an FFT is defined as a decision tree that has $m + 1$ exits, with one exit for each of the first $m - 1$ cues and two exits for the last cue [3].

An FFT relies on the so-called *non-compensatory* decision making; once the tree is exited, cues lower in the decision tree hierarchy cannot compensate for cue information higher in the hierarchy [1,11]. Surprisingly, by ignoring information, FFTs can be more accurate than statistical multivariate regression models ('less is more') [2]. This is because FFTs can be less susceptible to overfitting than the regression models [2].

Every cue in a FFT can correctly or incorrectly classify signal and noise. The performance of each cue can be described using SDT criteria [3]. Figure 1 illustrates the application of SDT to FFT. Thus, decision theories developed for small worlds can be directly linked to large-world decision theories [1,2]. The most important aspect of Fig. 1 is that the exit structure of the FFTs determines the ratio between false negatives and false positives. For example, the structure FFTyy on the far left side of Fig. 1 has a high hit rate (sensitivity) at the expense of a large rate of false alarms. The FFTnn on the far right side of the figure reflects the opposite: the FA is reduced at the expense of a large rate of false negatives.

EAT

Whereas SDT is based on Neyman–Pearson statistics, EAT is based on Abraham Wald's sequential decision theory, which assumes that decisions are made by accumulating evidence via a sequential sampling process [3]. Lee and Cummins showed [4] that sequential heuristics such as take the best (TTB) can be related to traditional 'rational' approaches to decision making by introducing an action threshold. TTB is a lexicographic heuristic for deciding between two objects, similar in spirit to an FFT, which assigns one object to one of two categories. Decisions in the framework of heuristic decision making often occur after the first piece of evidence is deemed sufficient for action (according to the stopping rule of TTB, similar to that of an FFT), whereas traditional rational approaches such as EUT require all of the available information to be sampled before a decision is made. Lee and Cummins [4] argued that 'by setting different threshold levels of evidence required for decision making, both the heuristics decision making and the rational models become special cases of a more general evidence accumulation account'.

Threshold model

A popular benefit–risk analysis in clinical medicine has relied on the application of the threshold model, which has been developed both in EUT and expected regret framework [6–8]. As noted earlier, in this paper we describe a signal in terms of a patient having a disease with some probability, *p*. According to the threshold model, when a physician is faced with uncertainty about whether to treat or simply further observe the patient, there must exist some *p* at which prescribing or not prescribing treatment are equal options (see Fig. 2) [6–8]. This probability is referred to as the threshold probability (*T*) [6–8]. Analysing the tree shown in Fig. 2, we find that the physician should prescribe treatment if *p* is

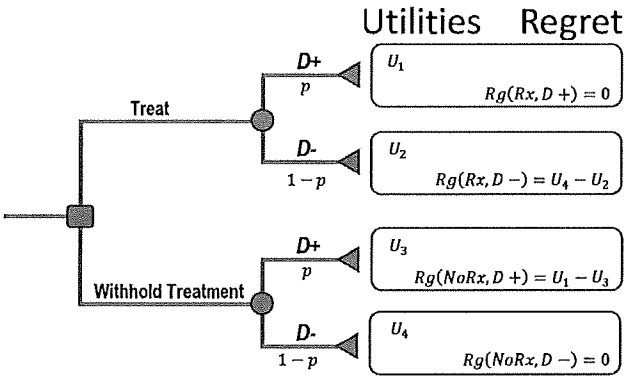


Figure 2 A decision tree from which the threshold model is derived. A physician has two choices: treat (e.g. prescribe statins) versus do not treat. The patient may or may not have disease (D) with the probability *p*. The probability at which a decision maker is indifferent between acting (e.g. treating) or not acting (e.g. withholding treatment) is called the threshold probability (see Appendix B for details).

larger than T and should withhold treatment if p is less than T. Figure 2 shows a decision tree where outcomes are expressed both in terms of utility and in terms of regret. Detailed derivations of the inequalities above are shown in Appendix B. Note that the regret model and EUT produce the same results as long as regret is a linear function of lost potential utilities, as is often the case in clinical settings [12,13].

One important aspect of the tree in Fig. 2 concerns the definition of treatment benefits and harms. In the original threshold model, these were defined as net benefits (B) and net harms (H). B represents the difference in the utility of the outcomes if a patient *with* disease was treated versus not treated; H is defined as the difference in the utility of the outcomes for a patient *without* the disease. That is, B = utility of true positives – utility of false negatives ($B = U_1 - U_3$ in Fig. 2); H = utility of true negatives – utility of false positives ($H = U_4 - U_2$ in Fig. 2) [6–8], similar to the classification of decision errors according to SDT.

This brief description of the threshold model lends itself to outlining a rationale for application of the threshold model to FFTs: a cue in an FFT can be selected according to the threshold model. Appendix B provides a brief derivation of the threshold model and Appendix C a side-by-side comparison of the threshold model with SDT.

Linking threshold model with SDT and FFT

As stated earlier, according to the threshold model, treatment is withheld if the posterior probability of disease is smaller than $\frac{1}{1+\frac{B}{H}}$ and is prescribed if the probability is larger than $\frac{1}{1+\frac{B}{H}}$.

Appendix C demonstrates that $\beta_{optimal}$ (Eq. 1) is equivalent to the action threshold in the threshold model.

Despite this mathematical equivalence between SDT and the threshold model, as demonstrated further below, the explicit linkage of the threshold to SDT (via EAT) can change the structure of an FFT. This, in turn, may affect both classification accuracy and decision consequences. More importantly, under some circumstances, using the threshold to calculate SDT statistics may not necessarily result in an orderly increase in c statistics from the left (FFTyy) to the right (FFTnn), as in the case of standard FFTs shown in Fig. 1. To distinguish it from a standard FFT, we henceforth refer to the threshold-derived FFT as an FFTT.

As noted, as long as the relationship remains linear, identical results can be derived using regret theory framework [12,13]. However, if preferences are elicited using regret scales [14], the differences between EUT and regret threshold model are often dramatic, even though their mathematical formulation is identical [8,15].

In the next three sections, we (1) show an application of the threshold model to both aid selection of cues for generating FFTs and help make better decisions; (2) describe the linking of the threshold model to EAT; and (3) provide a practical illustration of the approach outlined in this paper by focusing on decision differences between FFT and FFTT.

Selection of cues for FFT

Martignon and colleagues described two strategies to select cues for FFTs: MAX and ZIG (or zigzag, dual max) [16]. These strategies entail selecting a cue with the highest predictive accuracy but do not consider the consequences of a right or wrong decision.

The threshold model can be applied to determine when each cue should be considered *before* it is actually searched (Fig. 3).

The following relation holds for the situation before each cue is considered

1 Whenever the positive likelihood ratio ($LR+$) is *smaller* than $\beta_{optimal}$, we should not consider the cue (we should accept noise, or consider a new cue)

2 Whenever the negative likelihood ratio ($LR-$) is *larger* than $\beta_{optimal}$, we should not consider the cue (we should accept signal, or consider a new cue)

3 If both $(LR+) > \beta_{optimal}$ and $(LR-) < \beta_{optimal}$, we should proceed with considering a given cue and refine likelihood ratios accordingly.

Hence, the threshold model effectively determines the point when searching for more information will not change our decision. This is a critical aspect of effective application of heuristics: deciding when to stop further search, which, if continued, may prove to be counterproductive, costly or even detrimental [2].

It can be particularly challenging to determine a cue's cut-off when the cue is continuous. In such cases, the threshold model can be helpful. It can be shown that the optimal cut-off occurs at the point on a Receiver Operating Characteristics (ROC) curve plotting true positives (S) versus false positives (1 – C) where its slope is given by [17]:

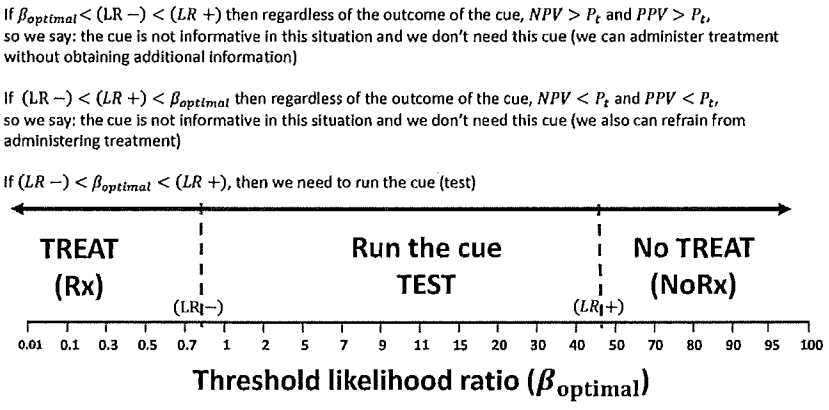


Figure 3 Selecting a cue according to the threshold model. Note how the threshold model effectively determines the point when search for more information will not change the decision. This is a critical aspect of effective application of heuristics: deciding when to stop further search, which, if continued, may turn to be counterproductive, costly or even detrimental (see text for details).

$$slope\ of\ ROC\ curve = \frac{H}{B} * \frac{p(Noise)}{p(Signal)} \quad (2)$$

Threshold model and EAT

As described earlier, EAT has been proposed as a unifying theory that can account for both heuristic decision making (such as TTB) and rational, EUT models [4]. Here we show how the threshold model can further unify SDT, FFT and EAT within EUT and the expected regret format. Lee and Cummins proposed that decision making follows sequential accrual of information to allow for comparison between two stimuli (choices) [4] and that when evidence exceeds a threshold, a decision is made. However, this threshold is never formally defined. Here we show that the threshold is equal to $\ln\beta_{optimal}$. The decision should be made when the (ln) of the sum of existing evidence $\geq\ln\beta_{optimal}$ according to

$$\sum_{a_i=1} \ln(LR_i+) + \sum_{a_i=0} \ln(LR_i-) \leq \ln\beta_{optimal} \quad (3)$$

In other words, for each patient with a sequence of cue answers (a_1, a_2, . . . , a_k) we can evaluate the sum of the logs of the likelihood ratios and see whether it is smaller than the log of $\beta_{optimal}$. If the left-hand sum is smaller than $\ln\beta_{optimal}$, treatment should be withheld, and if it is larger, treatment should be prescribed. Appendix D shows a formal linkage between EAT and the threshold model. As illustrated below, Eq. (3) allows treatment to be individualized for people with different cue values; it directly takes the consequences of treatment into account, whereas a standard FFT focuses on the accuracy of classification. That is, determining the accuracy statistics (d' and c) based on the combination of cues' true positives and true negatives in standard FFT–SDT format may not necessarily be equivalent to the classification based on the consequences of treatment according to Eq. (3) (see below for the specific illustration of the difference in the results between standard FFT and FFTT).

An illustration

We now illustrate linking these four theories – threshold, SDT, FFT and EAT – in the setting of prescribing statins for primary prevention of CVD. Statins are promoted as effective drugs for both primary and secondary prevention of CVD, the leading cause of mortality and morbidity in the United States and in most other economically developed nations [18,19]. Currently, the American College of Cardiology and the American Heart Association (ACC/AHA) recommend statins if the 10-year risk of myocardial infarction or stroke is ≥7.5% [20]. The risk of CVD can be estimated using multivariate regression models such as Framingham Risk Score (FRS) [21], derived from the Framingham Heart Study cohort [22]. However, many doctors find the use of the FRS and similar predictive models cumbersome and do not rely on them. In addition, the new guidelines recommend that doctors should not make their decisions based on laboratory values such as cholesterol levels [23]. As a consequence, most physicians rely on their clinical judgments and intuitive heuristics to advise their patients about taking statins. The important question is how formally defined heuristics such as FFT fare against multivariate regression models such as FRS. To answer this question, we obtained the data from the Framingham cohort from the National Heart, Lung, and Blood Institute in order to compare the accuracy of inferences based on FFT aided by SDT and threshold models with those based on the FRS model. Note, however, that our main goal was not to develop a new tool (FFT) that will replace decades of epidemiological and clinical research but instead to illustrate how new clinical tools can be developed using simple adaptive tools such as FFT while retaining a coherence with statistical decision theories.

The original data include the following eight variables: age, gender, total cholesterol, high-density cholesterol, systolic blood pressure, whether the patient received treatment for hypertension, whether the patient smoked and whether the patient had diabetes. All these variables were statistically associated with CVD over 10 years. Our interest was in generating an easy-to-use FFT containing ideally no more than three to five variables in order to retain its clinical usefulness. Given that new ACC/AHA guidelines de-emphasize laboratory measurements [20], we were particularly interested in generating an FFT based on clinical variables only. We aimed to compare the FFT with the FRS model at the recommended threshold of ≥7.5% of CVD [20].

To select a cue, we determined $\beta_{optimal}$ and compared it with the threshold LRs for each cue (Fig. 3). If the rule was not met, a cue was not selected. Table 1 shows the performance criteria of eight cues that comprise the FRS model in comparison with our selection criteria threshold rule shown in Fig. 3. The key determinants for the cue selection are benefits (B) and harms (H) of statins and the probability of CVD. The benefits and harms of statins in the setting of primary prevention of CVD are debatable. A Cochrane systematic review of four randomized trials enrolling 35 254 patients found a statistically significant effect of statins on two health outcomes only: statins reduced CVD by 1.33–2.5 percentage points in terms of absolute risk at the expense of an absolute increase of 0.4 percentage points in incidence of diabetes [19]. Other authors, however, estimated benefits of statins to be around 1% and incidence of diabetes about 1% [24]. In addition, statins have also been reportedly associated with a number of other harms (such as myalgia, liver test abnormalities, rhabdomyolysis) [25] that have not been precisely quantified in control trials [19] but that reduce the B/H ratio. Because of the uncertainty regarding the actual B/H of statins in this setting, there is an intense debate in the literature whether statins are more beneficial or more harmful to patients for primary prevention of CVD [23,26–28]. To take these uncertainties into account, we selected cues for the range of plausible B/H for primary prevention of CVD from as low as 0.5 to as high as 6. To make the FRS data more relevant, we assumed the baseline prevalence for CVD of 6% as the current best estimate of CVD in the general population [29].

From Table 1, it can be seen that across the wide range of B/H, three cues met our criteria for inclusion in an FFT: age, whether the patient has been diagnosed with diabetes and whether the patient is being treated for high blood pressure. Notably, these cues are always considered during a typical patient–physician encounter. Assuming that the B/H ratio of statins for primary prevention is between 1.33 and 2.5 [19], we determined the cut-off for age to be 39–49 years, with the best estimate at 44. Because our goal is to illustrate methods and not necessarily to generate a new tool for clinical practice, we report only analysis based on a B/H ratio of 2.5.

Table 1 Selection of cues to create a fast and frugal decision tree (FFT) for administration of statins in prevention of cardiovascular disease (CVD)

	Statins' benefit/harms (B/H) ratio*; pCVD = 0.25 and pCVD = 0.06 (bold)							
Cues	≤0.1[§]	0.2–0.6	0.7–1.1	1.2–1.6	1.7–2.1	**2.2–4**	4 to <5	≥5[¶]
Age (≥44)				**YES**	**YES**	**YES**	YES	
Gender						**YES**	YES	
Smoker								
Diabetic		YES	YES	**YES**	**YES**	**YES**		
Blood pressure treated			YES	**YES**	**YES**	**YES**		
Systolic blood pressure (≥125 mmHg)					YES	YES	YES	
Total cholesterol (≥250 mg dL^{-1})[†]					YES[‡]	YES	YES[‡]	
HDL cholesterol (≥40 mg dL^{-1})					YES	YES		

Bold texts refer to cues selected for the FFT described in this manuscript.

*At probability of cardiovascular disease (pCVD) of 6% (general population prevalence).

[†]If cut-off is selected at 220 mg dL^{-1} then total cholesterol can also be selected for B/H 1.2–1.6.

[‡]At cut-off of 235 mg dL^{-1}.

[§]No cue should be selected (completely uninformative) ($\beta_{optimal} > LR+$) (see Fig. 3).

[¶]No cues should be selected, that is, all patients should be treated ($\beta_{optimal} < LR-$) (see Fig. 3).

HDL, high-density lipoprotein.

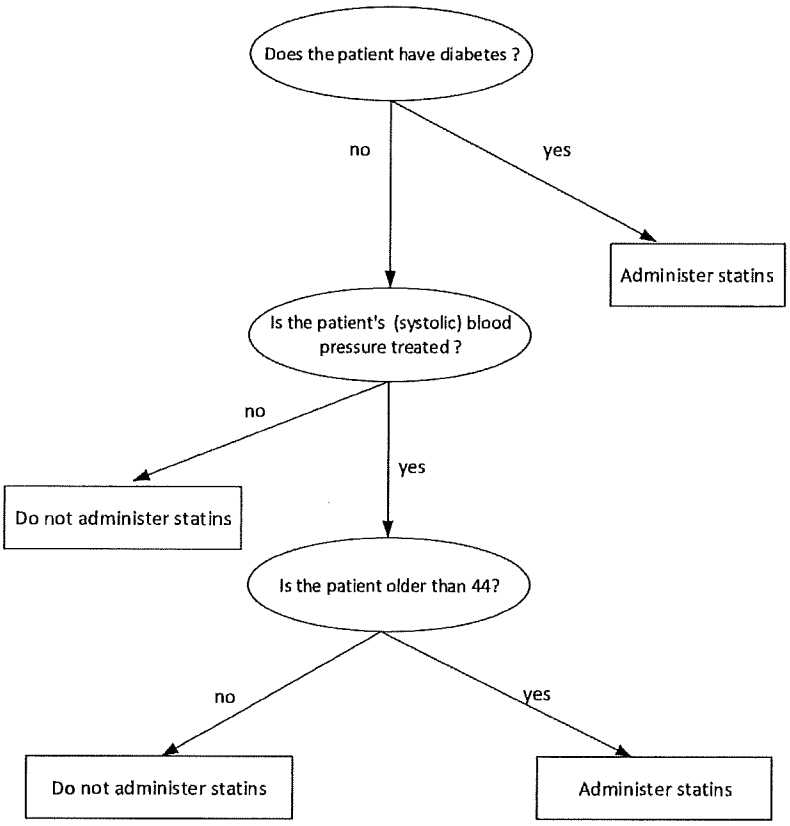


Figure 4 Fast-and-frugal tree (FFT) for prescribing statins for primary prevention of heart disease. An example of an FFTyn structure (see Fig. 1). The FFT was generated based on selection of three cues obtained from the initial patient visit: Does the patient have diabetes (yes/no)? Is the patient being treated for high pressure (yes/no)? Is the patient older than 44 (yes/no)? (see also Table 1). Note that because the last (in this case, third) cue has both exits, the FFT is identified by the cue exits before the final one (e.g. FFTyy, FFTyn, FFTny, FFTnn).

Figure 4 shows an FFT that we generated based on selection of these three cues obtained from the initial patient visit: Does the patient have diabetes (yes/no) (Y/N)? Is the patient being treated for high blood pressure (Y/N)? Is the patient older than 44 (Y/N)? The FFT stipulates that a physician will prescribe statins whenever the patient exits the tree structure after answering 'yes' and withhold prescribing statins whenever the patient exits the tree structure after answering 'no'. Therefore, there are four possible FFTs and for each individual patient there are eight possible exit paths: NNN, NNY, NYN, NYY, YNN, YNY, YYN, YYY (see also Figs 1 & 5).

The assumed B/H ratio (Table 1), however, gives equal weight to adverse effects of statins such as developing diabetes, muscle pain and liver toxicity. Because people may weigh these clinical

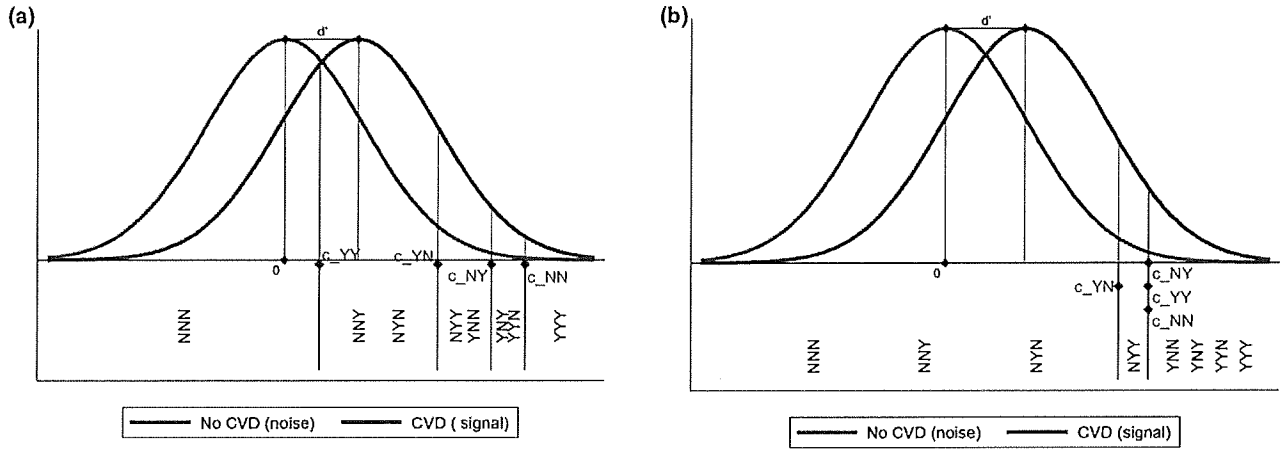


Figure 5 A relationship between fast-and-frugal trees (FFTs) with two cues and all possible paths that a patient can go through (denoted by a three-letter combination below the *x*-axis). Discriminability (*d'*) and decision criterion statistics (here denoted as *c*) are shown for each FFT first without considering the threshold (standard FFT – panel a) and where decision making depends on the threshold (FFTT – panel b). Treatment is indicated for all patients in the FFT paths to the right of the decision criterion (vertical lines passing through the green dots denote decision criterion *c*). Note, however, that in the case of FFTs (without a threshold) the signal (CVD = cardiovascular disease) significantly overlaps with noise (no CVD) for FFTyy, which means that the classification capacity of this FFT is likely unreliable. However, the structure of FFTyn, FFTny and FFnn shifts the decision criterion to the right with a much smaller overlap between the signal and noise. Here, all patients who exit on a 'yes' branch will receive treatment (denoted as the last letter 'y' in the FFT path), and all patients who exit on a 'no' branch will not receive treatment (denoted as the last letter 'n' in the FFT path). On the other hand, for FFTTs, the threshold (or in this case $\ln\beta_{optimal}$) can override the FFT exit decision, and treatment is sometimes withheld (or prescribed) for a patient who may have exited on a 'yes' (or 'no') exit. For example, as shown in panel (b), in an FFTT it is possible to treat patients along the path defined by FFTyn who would not be treated in an FFTyy. This is not possible in a standard FFT (a). (See Table 3 for comparison of treatment decisions between the FFT, FFTT and FRS models.)

outcomes differently, we also conducted a sensitivity analysis by assigning diverging weights to the failure to prevent CVD (regret of omission) and to incurring adverse events (regret of commission).

To test for the differences between the performance of FFTT (which applies the threshold model to derive classification and decision criteria) versus FFT (standard FFT, which refers to no threshold for making classification and decision recommendations) versus the FRS model (according to which statins should be prescribed if the probability of CVD is ≥7.5%), we also conducted a formal hypothesis test using the approach described by Wickens [30]. The null hypothesis of no difference was rejected if the probability of observing *z*-statistics was ≤0.05 [30].

Results

Table 2 shows the performance metrics of FFTT, FFT and the FRS model. Note that FRS performance characteristics remain identical, and that performance of FFFT does not always change predictably across all FFTT combination of cues. Unlike for FTT, where *c* statistics increase from FFTyy to FFTnn, *c* statistics are unpredictable for FTTT. That is because the threshold value is such that the balance of benefits and harms of statins remains constant across the combinations of cues and – depending on a given path – it is possible that more patients can be treated, for example, with FFTyn than with FFTyy (see Table 3 and Fig. 5). In general, FFFT retained high specificity across the cues, but at expense of very low sensitivity. The same held for all FFTs except for FFTyy, where specificity was more modest (80%) and sensitivity was higher, albeit not at a level to be considered informative (49%).

With respect to *d'* (discrimination, arguably the most important measure in SDT), FFTT and FRS performed better than FFT. Discriminability of FFTTyy was superior to FFTyy ($P = 0.0078$), whereas no difference was detected between FFTTyy and the FRS model ($P = 0.61$); the FRS model was, however, superior to FFT ($P = 0.003$). At the same time, no difference was detected between FFTTnn versus FFTnn and the FRS model ($P = 0.412$ for FFTTnn versus FFTnn; $P = 0.60$ for FFTnn versus FRS; $P = 0.76$ for FFTTnn versus FRS).

Overall, the performances of both FFT and the FRS model were far from perfect. This is not particularly surprising in light of the recent external validation of four most popular CVD risk prediction models (including FRS) demonstrating that all existing models fall short of being highly predictive of CVD [31,32]. All the models tested had high and positive values for *c*. The *c* value for FFTTyy was higher than for FFTyy ($P = 0.00001$) and FRS ($P = 0.00001$) but was identical for FFTTyn and FFTyn (both had a larger *c* value than the FRS model, at $P = 0.00001$). For FFTny and FFTnn, the *c* value was larger than that of their respective FFTTs (at $P = 0.00001$). Both *c* values remain larger than that of FRS ($P = 0.00001$).

In general, a positive *c* value means that more weight is placed on avoiding false positives, while a negative *c* value indicates a preference for avoiding not treating someone with CVD (i.e. avoiding false negatives). In light of the current controversy about B/H of statins and the estimation that more than a billion people worldwide would be prescribed the treatment if the AHA/CCA guidelines were adhered to [28], it is important to identify a simple tool for avoiding unnecessary treatment. That is, more important than the accuracy statistics shown in Table 2 are the decision

Table 2 Performance of FFTT versus FFT versus FRS* model[†§]

	FFT (YY) with threshold	FFT (YY) w/o threshold	FRS (EUT = 0.075)[‡] model
Sensitivity	0.0197 (0.0196, 0.0197)	0.4934 (0.4934, 0.4935)	0.4348 (0.4347, 0.4348)
Specificity	0.9989 (0.9989, 0.9989)	0.8074 (0.8073, 0.8075)	0.8856 (0.8856, 0.8857)
LR+	18.3996 (5.4321, 62.3226)	2.5621 (2.3178, 2.8321)	3.8013 (3.3502, 4.3132)
LR–	0.9814 (0.9724, 0.9905)	0.6274 (0.5871, 0.6705)	0.6382 (0.6021, 0.6765)
c	2.5658 (2.5624, 2.5692)	0.4424 (0.4416, 0.4432)	0.6839 (0.6832, 0.6846)
d'	1.0096 (1.0033, 1.0159)	0.8520 (0.8471, 0.8568)	1.0394 (1.0349, 1.0438)
	FFT (YN) with threshold	**FFT (YN) w/o threshold**	**FRS (EUT = 0.075) model**
Sensitivity	0.0677 (0.0676, 0.0678)	0.0677 (0.0676, 0.0678)	0.4348 (0.4346, 0.4349)
Specificity	0.9915 (0.9915, 0.9915)	0.9915 (0.9915, 0.9915)	0.8856 (0.8856, 0.8857)
LR+	7.9220 (4.9748, 12.6153)	7.9220 (4.9748, 12.6153)	3.8013 (3.3502, 4.3132)
LR–	0.9403 (0.9338, 0.9572)	0.9403 (0.9238, 0.9572)	0.6382 (0.6021, 0.6765)
c	1.9390 (1.9374, 1.9406)	1.9390 (1.9374, 1.9406)	0.6839 (0.6832, 0.6846)
d'	0.8916 (0.8874, 0.8957)	0.8916 (0.8874, 0.8958)	1.0394 (1.0349, 1.0438)
	FFT (NY) with threshold	**FFT (NY) w/o threshold**	**FRS (EUT = 0.075) model**
Sensitivity	0.0196 (0.0195, 0.0196)	0.0152 (0.0152, 0.0153)	0.4348 (0.4346, 0.4349)
Specificity	0.9989 (0.9989, 0.9989)	0.9989 (0.9989, 0.9989)	0.8856 (0.8856, 0.8857)
LR+	18.4174 (5.4373, 62.3837)	14.3246 (4.1257, 49.7360)	3.8013 (3.3502, 4.3132)
LR–	0.9815 (0.9725, 0.9906)	0.9858 (0.9779, 0.9953)	0.6382 (0.6021, 0.6765)
c	2.5675 (2.5641, 2.5709)	2.6183 (2.6139, 2.6227)	0.6839 (0.6832, 0.6846)
d'	1.0094 (1.0031, 1.0157)	0.9078 (0.8997, 0.9160)	1.0394 (1.0349, 1.0438)
	FFT (NN) with threshold	**FFT (NN) w/o threshold**	**FRS (EUT = 0.075) model**
Sensitivity	0.0196 (0.0195, 0.0196)	0.0060 (0.0059, 0.0060)	0.4348 (0.4346, 0.4349)
Specificity	0.9989 (0.9989, 0.9989)	0.9998 (0.9998, 0.9998)	0.8856 (0.8856, 0.8857)
LR+	18.4174 (5.4373, 62.3837)	33.7405 (1.8674, 609.6319)	3.8013 (3.3502, 4.3132)
LR–	0.9815 (0.9725, 0.9906)	0.9942 (0.9892, 0.9992)	0.6382 (0.6021, 0.6765)
c	2.5675 (2.5641, 2.5709)	3.0430 (3.0349, 3.0511)	0.6839 (0.6832, 0.6846)
d'	1.0094 (1.0031, 1.0157)	1.0584 (1.0462, 1.0706)	1.0394 (1.0349, 1.0438)

*FFT only refers to two cues as the last cue has both exits. This FFT asks if the patient has diabetes (yes/no) (Y/N), is on blood pressure medication (BPRx) (Y/N) and is older than 44 (Y/N) (see Fig. 4).

[†]Assumes benefit/harms of statins for primary prevention of cardiovascular disease (CVD) of 2.5 and probability (pCVD) = 0.06; FFTT – fast and frugal tree that takes threshold in consideration; FFT – standard fast and frugal tree that takes no threshold into consideration (w/o); FRS – Framingham Risk Score.

[‡]ACC/AHA – The American College of Cardiology and American Heart Association recommends statins if pCVD ≥7.5% (see text for details).

[§]Note that because in both FFTT_ny and FFTT_nn trees 99.44% of the patients take the first exit, FFTT_ny and FFTT_nn result in almost identical results.

consequences of FFTT versus FFT versus FRS in terms of prescribing statins to people who actually end up developing CVD versus not prescribing statins to those who do not develop it.

Table 3 shows a comparison of treatments according to FFTT, FFT and AHA/ACC guidelines. In general, there is a statistically significant association – ranging from small to moderate – between FFT paths and risks for CVD. Note, however, the wide ranges in the probability estimates of CVD according to the FRS for each combination of the cues, which confirms the previously mentioned unreliability of the FRS estimates [31,32]. All models concur that patients who do not have diabetes, are not being treated for high blood pressure and are younger than 44 should not be prescribed statins ('NNN'). Likewise, all models agree that patients who have diabetes (irrespective of treatment for hypertension and age) should also be recommended statins. The models differ in their recommendations for patients with other risk profiles, but given the absence of universally agreed prediction and recommendations standards, it is not surprising that no clear-cut guideline for prescribing statins has emerged for all patients [23,26–28]. Indeed, the tendency to avoid unnecessary treatment becomes further apparent when the B/H ratio decreases: the *c* values become even more positive, indicating that our FFT/FFTT place more value on avoiding unnecessary treatment (i.e. false positives; results not shown). Of course, if our goal is to avoid false negatives, this FFT/FFTT performs only modestly, but so does FRS.

This consideration assumes that we weigh consequences of false negatives versus false positives equally. What would happen if a decision maker weighed these differently [33]? When the effect of differential weighting of false negatives (regret of omission) versus false positives (regret of commission) on the discriminatory capability of our FFT was assessed, avoidance of false positives was associated with both larger *d'* and larger *c* (Fig. 6). As Fig. 6 shows, under some circumstances our FFT (when weighted by expected consequences) can even achieve virtually perfect discrimination ($d' \geq 4$).

Table 3 Who should be treated with statins? Comparison of individualized decision making according to FFTT/evidence accumulation theory versus FFT versus FRS model/ACC/AHA guidelines

(a) FFTyy

FFTyy				Treatment		
Possible paths for this FFT†	pCVD* (median and range)	logLRsum**	Number of patients	According to FFTT*†	According to FTT*†	According to FRS model (ACC/AHA guidelines)*†
NNN	1.7% (0.43–24.5%)	−0.514	2732	No	No	No
NNY	7.45% (1.1–47.1%)	0.873	871	No	Yes	No
NY	11.9% (1.18–62.4%)	1.438	101	No	Yes	Yes
Y	21.7% (6.5–57.3%)	2.913	21	Yes	Yes	Yes
			3725			

*pCVD – probability of cardiovascular disease according to FRS (because highly skewed data, summary statistics expressed as median and range); FRS – Framingham Risk Score; ACC/AHA – The American College of Cardiology and American Heart Association recommends statins if pCVD ≥ 7.5%; FFTT – fast-and-frugal tree that takes threshold in consideration (according to the criterion stated in **); FFT – standard fast-and-frugal tree that takes no threshold into consideration but according to which each patient whose path ends with 'Y' should be treated (see text for details).

** $\sum_{a_i=1}\ln(LR_i+)+\sum_{a_i=0}\ln(LR_i-)$; treat if logLRsum $>ln\beta_{optimal}$ *(1.83; based on benefit/harms of statins of 2.5 and pCVD = 0.06)*; otherwise withhold treatment.

†Spearman rho correlation coefficient between pCVD and the FFT path: 0.64; $P < 0.000001$. Spearman rho correlation coefficient between pCVD and treatment according to: FFTT (0.11; $P < 0.000001$); FFT (0.60; $P < 0.000001$) and FRS/ACC/AHA guidelines (0.68; $P < 0.000001$).

(b) FFTyn

FFTyn				Treatment		
Possible paths for this FFT†	pCVD* (median and range)	logLRsum**	Number of patients	According to FFTT*†	According to FTT*†	According to FRS model (ACC/AHA guidelines)*†
NN	2.4% (0.43–47.1%)	−0.073	3603	No	No	No
NYN	6.4% (1.1–24.7%)	0.997	36	No	No	No
NYY	15.9% (2.5–62.4%)	2.38	65	Yes	Yes	Yes
Y	21.7% (6.5–57.3%)	2.913	21	Yes	Yes	Yes
			3725			

*pCVD – probability of cardiovascular disease according to FRS (because highly skewed data, summary statistics expressed as median and range); FRS – Framingham Risk Score; ACC/AHA – The American College of Cardiology and American Heart Association recommends statins if pCVD ≥ 7.5%; FFTT – fast-and-frugal tree that takes threshold in consideration (according to the criterion stated in **); FFT – standard fast-and-frugal tree that takes no threshold into consideration but according to which each patient whose path ends with 'Y' should be treated (see text for details).

** $\sum_{a_i=1}\ln(LR_i+)+\sum_{a_i=0}\ln(LR_i-)$; treat if logLRsum $>ln\beta_{optimal}$ *(1.83; based on benefit/harms of statins of 2.5 and pCVD = 0.06)*; otherwise withhold treatment.

†Spearman rho correlation coefficient between pCVD and the FFT path: 0.232; $P < 0.000001$. Spearman rho correlation coefficient between pCVD and treatment according to FFTT (0.223; $P < 0.000001$), FFT (0.223; $P < 0.000001$) and FRS/ACC/AHA guidelines (0.68; $P < 0.000001$).

(c) FFTny

FFTny				Treatment		
Possible paths for this FFT†	pCVD* (median and range)	logLRsum**	Number of patients	According to FFTT*†	According to FTT*†	According to FRS model (ACC/AHA guidelines)*†
N	2.5% (0.43–62.4%)	−0.018	3723	No	No	No
YNN	12.7% (8.9–15.9%)	2.41	4	Yes	No	Yes
YNY	24.9% (10.4–57.3%)	3.8	10	Yes	Yes	Yes
YY	35.4% (6.5–46.3%)	4.37	7	Yes	Yes	Yes
			3744			

*pCVD – probability of cardiovascular disease according to FRS (because highly skewed data, summary statistics expressed as median and range); FRS – Framingham Risk Score; ACC/AHA – The American College of Cardiology and American Heart Association recommends statins if pCVD ≥7.5%; FFTT – fast-and-frugal tree that takes threshold in consideration (according to the criterion stated in **); FFT – standard fast-and-frugal tree that takes no threshold into consideration but according to which each patient whose path ends with 'Y' should be treated (see text for details).

** $\sum_{a_i=1}\ln(LR_i+)+\sum_{a_i=0}\ln(LR_i-)$; treat if logLRsum $>ln\beta_{optimal}$ *(1.83; based on benefit/harms of statins of 2.5 and pCVD = 0.06)*; otherwise withhold treatment.

†Spearman rho correlation coefficient between pCVD and the FFT path: 0.115; $P < 0.000001$. Spearman rho correlation coefficient between pCVD and treatment according to FFTT (0.11; $P < 0.000001$), FFT (0.10; $P < 0.000001$) and FRS/ACC/AHA guidelines (0.68; $P < 0.000001$).

Table 3 *Continued*

(d) FFTnn						
FFTnn				Treatment		
Possible paths for this FFT†	pCVD* (median and range)	logLRsum**	Number of patients	According to FFTT*†	According to FTT*†	According to FRS model (ACC/AHA guidelines)*†
N	2.5% (0.43–62.4%)	−0.018	3723	No	No	No
YN	18% (8.9–15.9%)	2.85	14	Yes	No	Yes
YYN	7.7% (6.5–8.8%)	3.92	2	Yes	No	Yes
YYY	35.5% (22.1–46.3%)	5.31	5	Yes	Yes	Yes
			3744			

*pCVD – probability of cardiovascular disease according to FRS (because highly skewed data, summary statistics expressed as median and range); FRS – Framingham Risk Score; ACC/AHA – The American College of Cardiology and American Heart Association recommends statins if pCVD ≥7.5%; FFTT – fast-and-frugal tree that takes threshold in consideration (according to the criterion stated in **); FFT – standard fast-and-frugal tree that takes no threshold into consideration but according to which each patient whose path ends with 'Y' should be treated (see text for details).

** $\sum_{a_i=1}\ln(LR_i+)+\sum_{a_i=0}\ln(LR_i-)$; treat if logLRsum $>ln\beta_{optimal}$ *(1.83; based on benefit/harms of statins of 2.5 and pCVD = 0.06)*; otherwise withhold treatment.

†Spearman rho correlation coefficient between pCVD and the FFT path: 0.115; $P < 0.000001$. Spearman rho correlation coefficient between pCVD and treatment according to FFTT (0.11; $P < 0.000001$), FFT (0.06; $P = 0.0008$) and FRS/ACC/AHA guidelines (0.68; $P < 0.000001$).

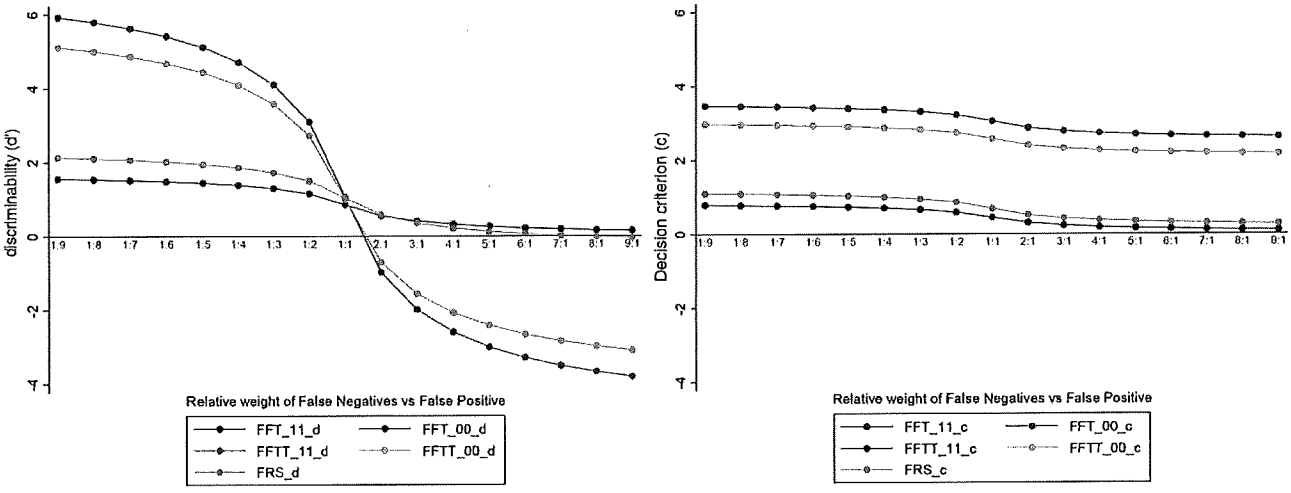


Figure 6 Effect of differential weighting of false negatives (regret of omission) versus false positives (regret of commission) on FFTs' discriminability (d') and decision criterion (c) (for pCVD = 0.06 and B/H = 2.5). The effect on FFTyy, FFTTyy, FFTnn and FFTTnn is shown. Note how weighting may affect signal theory statistics to approach perfect discrimination under some circumstances ($d' > 4$).

Discussion

In this paper, we show how a simple threshold model [34] can function as a foundational building block that can serve as a link between SDT, FFTs and EAT. We draw attention to two issues of relevance for decision sciences. First, many seemingly different concepts used in different disciplines are actually equivalent. Second, by making this connection, we arrived at what we believe are novel findings. For one, Eq. (3) explicitly links the threshold model to EAT and FFT, thereby enabling creation of decision criteria that take into account both the accuracy statistics of FFT and the consequences built in the threshold model. As a result, decision recommendations based on standard FFT often differ from FFT that take threshold into consideration (FFTT), as demonstrated in this paper (see Table 3 and Fig. 5). Another novel result is that threshold criteria can be used as a strategy for selecting cues to construct an FFT (Table 1). Finally, although the threshold model describes a compensatory decision-making strategy and the FFT represents a non-compensatory approach to decision making, we show that both strategies are ultimately linked within a broader theoretical framework.

Of key interest, we believe that we described a new method for creating and evaluating FFTs of relevance to medical decision making. Clinical decision making is dominated by heuristic thinking and by simple decision trees resembling FFTs. Yet few methods have actually been developed to show how these FFTs can be constructed, evaluated and applied.

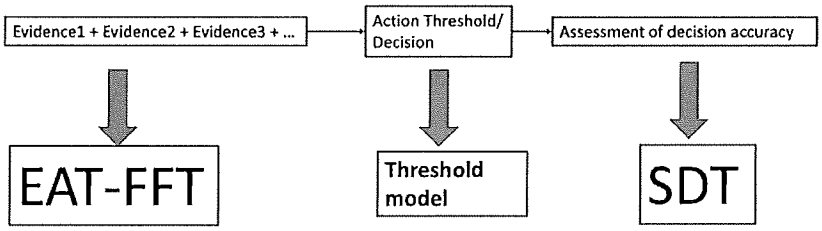


Figure 7 Model of decision making based on integration of evidenced accumulation theory (EAT), fast-and-frugal trees (FFT), threshold model and signal detection theory (SDT). FFT–EAT proposes that evidence is accumulated sequentially after which decision is made when certain threshold is exceeded (see Eq. 3). Such a decision is inevitably associated with false-positive versus false-negative errors, which are best appreciated within SDT framework (see Figs 1, 5 & 6).

Although the purpose of this paper is methodological, it is interesting to note that we succeeded in generating an FFT of practical importance. Consistent with FFT principles, we showed that our simple FFT, constructed of readily available cues, indicates that the patients who do not have diabetes and/or do not require antihypertensives can avoid statins. Likewise, we showed that patients who have diabetes should be treated. Both recommendations agree with the complex FRS model and AHA/ACC guidelines. In addition, it is instructive to note that our simple FFT departs from the AHA/ACA guidelines reflecting the disagreement in the field as to who should or should not be treated with statins [23,26–28,35]. This raises another interesting application of our proposed methodology: When an FFT does not agree with currently employed risk models and guideline recommendations (Table 3), this likely identifies open questions that need to be settled in future research.

In the final analysis, we showed how decisions depend on reliable information [36] – the results are highly sensitive to the estimate of the benefits and harms of interventions. Moreover, efficient cognitive processes require that we ignore part of the information available, particularly information that is unreliable, increases the estimation error or is costly to obtain [1,2]. Selection of a particular cue – the cue that may end up being used for constructing a simple FFT – should be evidence based and rely on as accurate information as possible. Specifically, we defined the criteria that can be used for selection of cues for FFT and under which circumstances such a selection would not be helpful.

In addition, we showed that the threshold model can provide a link between EUT and expected regret theory and SDT, FFT and EAT, thereby enabling us to integrate several theories of decision making under one conceptual umbrella [5]. By contrasting the logarithm of the sum of the accumulated evidence for positive (LR+) and negative (LR–) values for each cue with the action threshold, we define the decision criteria (for giving treatment). Most importantly, the latter offers an attractive possibility of individualizing treatment according to the combination of cues that a patient has.

Our method is not without limitations, however. Equation (3) assumes independence of cues, which may or may not be correct. Yet, as the bias–variance dilemma implies, assuming independence is likely to increase the bias of the model but simultaneously decrease estimation error (variance). Thus, the real question is one of the trade-off between the two sources of error: bias versus variance [2]. Future work should improve on the proposed methodology by taking into account this trade-off in the selection of cues.

Figure 7 shows a summary interpretation of our approach to unify several theoretical concepts. FFT–EAT proposes that evidence is accumulated sequentially after which decision is made when certain threshold is exceeded. Such a decision is inevitably associated with false-positive versus false-negative errors, which are best appreciated within SDT framework.

In conclusion, we showed that theories that at first glance appeared to be disconnected can be effectively linked under the theory integration programme [5]. Connecting apparently unrelated theories in different disciplines likely leads to discovery of new relationships. We believe that the programme of analysing the conceptual links between theoretical concepts can enable researchers from different disciplines to relate more fruitfully to each other.

Acknowledgement

This research is supported in part by DOD grant (#W81 XWH 09-2-0175) PI: Dr. Djulbegovic.

References

1. Gigerenzer, G. & Gaissmaier, W. (2011) Heuristic decision making. *Annual Review of Psychology*, 62, 451–482.
2. Gigerenzer, G. & Brighton, H. (2009) Homo heuristicus: why biased minds make better inferences. *Topics in Cognitive Science*, 1 (1), 107–143.
3. Luan, S., Schooler, L. J. & Gigerenzer, G. (2011) A signal-detection analysis of fast-and-frugal trees. *Psychological Review*, 118 (2), 316–338.
4. Lee, M. D. & Cummins, T. D. R. (2004) Evidence accumulation in decision making: unifying 'take the best' and 'rational' models. *Psychonomic Bulletin & Review*, 11 (2), 343–352.
5. Gigerenzer, G. (2015) A theory integration program. *Decision* (in press).
6. Pauker, S. G. & Kassirer, J. (1980) The threshold approach to clinical decision making. *The New England Journal of Medicine*, 302, 1109–1117.
7. Pauker, S. G. & Kassirer, J. P. (1975) Therapeutic decision making: a cost benefit analysis. *The New England Journal of Medicine*, 293, 229–234.
8. Mickes, L., Wixted, J. T. & Wais, P. E. (2007) A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review.*, 14 (5), 858–865.

9. Stanislaw, H. & Todorov, N. (1999) Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31 (1), 137–149.
10. Gigerenzer, G., Hertwig, R. & Pachur, T. (eds) (2011) Heuristics. The Foundation of Adaptive Behavior. New York: Oxford University Press.
11. Jenny, M. A., Pachur, T., Williams, S. L., Becker, E. & Margraf, J. (2013) Simple rules for detecting depression. *Journal of Applied Research in Memory and Cognition*, 2, 149–157.
12. Hozo, I. & Djulbegovic, B. (2008) When is diagnostic testing inappropriate or irrational? Acceptable regret approach. *Medical Decision Making*, 28 (4), 540–553.
13. Hozo, I. & Djulbegovic, B. (2009) Clarification and corrections of acceptable regret model. *Medical Decision Making*, 29, 323–324.
14. Tsalatsanis, A., Hozo, I., Vickers, A. & Djulbegovic, B. (2010) A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making. *BMC Medical Informatics and Decision Making*, 10 (1), 51.
15. Djulbegovic, B., Elqayam, S., Reljic, T., *et al.* (2014) How do physicians decide to treat: an empirical evaluation of the threshold model. *BMC Medical Informatics and Decision Making*, 14 (1), 47.
16. Martignon, L., Katsikopoulos, K. V. & Woikeb, J. K. (2008) Categorization with limited resources: a family of simple heuristics. *Journal of Mathematical Psychology*, 52, 352–361.
17. Sox, H. C., Higgins, M. C. & Owens, D. (2013) Medical Decision Making, 2nd edn. Chicester: Wiley-Blackwell.
18. Cholesterol Treatment Trialists C (2012) The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. *The Lancet*, 380 (9841), 581–590.
19. Taylor, F., Huffman, M. D., Macedo, A. F., *et al.* (2013) Statins for the primary prevention of cardiovascular disease. *The Cochrane Database of Systematic Reviews*, (1), CD004816.
20. Stone, N. J., Robinson, J. G., Lichtenstein, A. H., *et al.* (2014) 2013 ACC/AHA Guideline on the Treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines. *Journal of the American College of Cardiology*, 63 (25 Pt B), 2889–2934.
21. D'Agostino, R. B. Sr, Vasan, R. S., Pencina, M. J., *et al.* (2008) General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*, 117 (6), 743–753.
22. Framingham Heart Study. (2015) Framingham Heart Study. Available at: https://www.framinghamheartstudy.org/.
23. Pencina, M. J., Navar-Boggan, A. M., D'Agostino, R. B. Sr, *et al.* (2014) Application of new cholesterol guidelines to a population-based sample. *The New England Journal of Medicine*, 370 (15), 1422–1431.
24. Newman, D. (2015) Statin drugs given for 5 years for heart disease prevention (without known heart disease). Available at: http://www.thennt.com/nnt/statins-for-heart-disease-prevention-without-prior-heart-disease/ (last accessed 4 September 2015).
25. Newman, C. B. & Tobert, J. A. (2015) Statin intolerance: reconciling clinical trials and clinical experience. *JAMA: The Journal of the American Medical Association*, 313 (10), 1011–1012.
26. Redberg, R. F. & Katz, M. H. (2012) Healthy men should not take statins. *JAMA: The Journal of the American Medical Association*, 307 (14), 1491–1492.
27. Ioannidis, J. P. (2014) Guidelines for cardiovascular risk assessment and cholesterol treatment – reply. *JAMA: The Journal of the American Medical Association*, 311 (21), 2235–2236.
28. Ioannidis, J. P. (2014) More than a billion people taking statins? Potential implications of the new cardiovascular guidelines. *JAMA: The Journal of the American Medical Association*, 311 (5), 463–464.
29. Center for Disease and Control (2010) Prevalence of Coronary Heart Disease – United States, 2006–2010.
30. Wickens, T. D. (2002) Elementary Signal Detection Theory. Oxford: Oxford University Press.
31. DeFilippis, A. P., Young, R., Carrubba, C. J., *et al.* (2015) An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort: calibration and discrimination among CVD risk scores. *Annals of Internal Medicine*, 162 (4), 266–275.
32. Ridker, P. M. & Cook, N. R. (2015) Comparing cardiovascular risk prediction scores. *Annals of Internal Medicine*, 162 (4), 313–314.
33. Zeelenberg, M. & Pieters, R. (2007) A theory of regret regulation 1.0. *Journal of Consumer Psychology: The Official Journal of the Society for Consumer Psychology*, 17, 3–18.
34. Djulbegovic, B., van den Ende, J., Hamm, R. M., *et al.* (2015) When is rational to order a diagnostic test, or prescribe treatment: the threshold model as an explanation of practice variation. *European Journal of Clinical Investigation*, 45 (5), 485–493.
35. Montori, V. M., Brito, J. P. & Ting, H. H. (2014) Patient-centered and practical application of new high cholesterol guidelines to prevent cardiovascular disease. *JAMA: The Journal of the American Medical Association*, 311 (5), 465–466.
36. Djulbegovic, B., Guyatt, G. H. & Ashcroft, R. E. (2009) Epistemologic inquiries in evidence-based medicine. *Cancer Control: Journal of the Moffitt Cancer Center*, 16 (2), 158–168.
37. Djulbegovic, B. & Hozo, I. (2007) When should potentially false research findings be considered acceptable? *PLoS Medicine*, 4 (2), e26.
38. Hozo, I. & Djulbegovic, B. (2009) Will insistence on practicing medicine according to expected utility theory lead to an increase in diagnostic testing? *Medical Decision Making*, 29, 320–322.

HEPATOLOGY
Official Journal of the American Association for the Study of Liver Diseases

When to Perform Hepatic Resection for Intermediate-Stage Hepatocellular Carcinoma

Alessandro Cucchetti,[1] Benjamin Djulbegovic,[2] Athanasios Tsalatsanis,[2] Alessandro Vitale,[3] Iztok Hozo,[4] Fabio Piscaglia,[1] Matteo Cescon,[1] Giorgio Ercolani,[1] Francesco Tuci,[3] Umberto Cillo,[3] and Antonio Daniele Pinna[1]

Transcatheter arterial chemoembolization (TACE) is the first-line therapy recommended for patients with intermediate hepatocellular carcinoma (HCC). However, in clinical practice, these patients are often referred to surgical teams to be evaluated for hepatectomy. After making a treatment decision (e.g., TACE or surgery), physicians may discover that the alternative treatment would have been preferable, which may bring a sense of regret. Under this premise, it is postulated that the optimal decision will be the one associated with the least amount of regret. Regret-based decision curve analysis (Regret-DCA) was performed on a Cox's regression model developed on 247 patients with cirrhosis resected for intermediate HCC. Physician preferences on surgery versus TACE were elicited in terms of regret; threshold probabilities (P_t) were calculated to identify the probability of survival for which physicians are uncertain of whether or not to perform a surgery. A survey among surgeons and hepatologists regarding three hypothetical clinical cases of intermediate HCC was performed to assess treatment preference domains. The 3- and 5-year overall survival rates after hepatectomy were 48.7% and 33.8%, respectively. Child-Pugh score, tumor number, and esophageal varices were independent predictors of survival ($P < 0.05$). Regret-DCA showed that for physicians with P_t values of 3-year survival between 35% and 70%, the optimal strategy is to rely on the prediction model; for physicians with Pt <35%, surgery should be offered to all patients; and for P_t values >70%, the least regretful strategy is to perform TACE on all patients. The survey showed a significant separation among physicians' preferences, indicating that surgeons and hepatologists can uniformly act according to the regret threshold model. *Conclusion:* Regret theory provides a new perspective for treatment-related decisions applicable to the setting of intermediate HCC. (HEPATOLOGY 2015;61:905-914)

Hepatocellular carcinoma (HCC) is one of the most common cancers worldwide. It is the third leading cause of cancer-related death and is usually associated with cirrhosis.[1,2] The treatment of HCC is complex and different from many other cancers because of the conflicting needs to be oncologically radical, while, at the same time, aiming to preserve liver function, because the patient may die either from cancer or from progression of cirrhosis. Hepatic resection, even though radical, has usually only a minor role in the treatment of multiple, large HCC.[2-4] Transcatheter arterial chemoembolization (TACE) is the recommended first-line treatment for such a tumor stage (intermediate Barcelona Clinic Liver Cancer [BCLC] stage).[2,3] The main reason for the narrow indications for hepatic resection (HR) is the unfavorable outcomes expected in patients having an intermediate HCC.[2-5] However, in clinical practice,

Abbreviations: BCLC, Barcelona Clinic Liver Cancer; CI, confidence interval; CT, computed tomography; DCA, decision curve analysis; DVAS, dual visual analog scale; EV, esophageal varices; FN, false negative; FP, false positive; HCC, hepatocellular carcinoma; HR, hepatic resection; INR, international normalized ratio; IQR, interquartile range; MELD, Model for End-Stage Liver Disease; OS, overall survival; P_t, threshold probability; TACE, transcatheter arterial chemoembolization; US, ultrasound.

From the [1]Department of Medical and Surgical Sciences, S. Orsola-Malpighi Hospital, Alma Mater Studiorum–University of Bologna, Bologna, Italy; [2]Division of Evidence-based Medicine, Department of Internal Medicine, University of South Florida and H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL; [3]Department of General Surgery and Organ Transplantation, Hepatobiliary Surgery and Liver Transplant Unit, University of Padua, Padua, Italy; and [4]Department of Mathematics and Actuarial Science, Indiana University Northwest, Gary, IN.

Received December 21, 2013; accepted July 10, 2014.

such patients are frequently referred to surgical teams for evaluation and treatment.[5-7] This deviation from the guideline recommendations[2,3] is the consequence of the uncertainty regarding the optimal treatment for intermediate-stage HCCs, the balance between therapeutic efficacy and preservation of liver function, and the lack of uniformity of stage definition resulting in the inclusion of a heterogeneous patient population.[5-8]

Both hepatologists and surgeons are daily involved in the decision of whether or not to operate on a patient with intermediate-stage HCC. The decision is fraught with uncertainty and consequences that are impossible to predict at the time of the decision. After making a decision under such uncertainty, the physician may discover, when learning the relevant outcomes, that an alternative approach would have been preferable. This knowledge may bring a sense of loss or regret. Regret theory postulates that choices may be influenced by the decision maker's anticipation that certain outcomes will be associated with high regret, which he or she would like to avoid or minimize.[9-11] Thus, if we seek to minimize regret, the optimal choice would be the one associated with the least amount of regret. In general, regret can be felt as a result of wrong action (regret of commission) or failure to act (regrer of omission).[12] The assessment of regret of commission versus omission can be used to compute a threshold value at which the physician is uncertain about which treatment strategy to adopt.[13-15] The threshold value then can be employed to facilitate decision making through the technique known as regret–decision curve analysis (Regret-DCA).[14-17] In the present study, we assessed whether decisions regarding performing surgery or TACE on a patient with intermediate HCC can be aided by a prognostic model, depending on regret-based thresholds. We also recorded a distribution of the threshold probabilities to assess whether the preferences vary or are clustered within a relatively narrow range, which has important implications from the decision-making point of view.

Materials and Methods

Description of the Regret Model. The regret model assumes that there must be some probability of an event at which regret of omission (e.g., failure to perform resection in a patient with HCC) is equal to regret of commission (e.g., unnecessary surgery).[11,13-15] Similarly to classic expected-utility–based decision theory,[16,17] the regret model assumes that a patient should be treated if the probability of an event is equal or greater than a specific threshold probability.[14-16] The regret model was applied here to estimate the threshold probability at which the decision maker is indifferent between HR versus withholding surgery and administer TACE for intermediate HCC. The patient should be referred to surgery if his or her probability of survival is

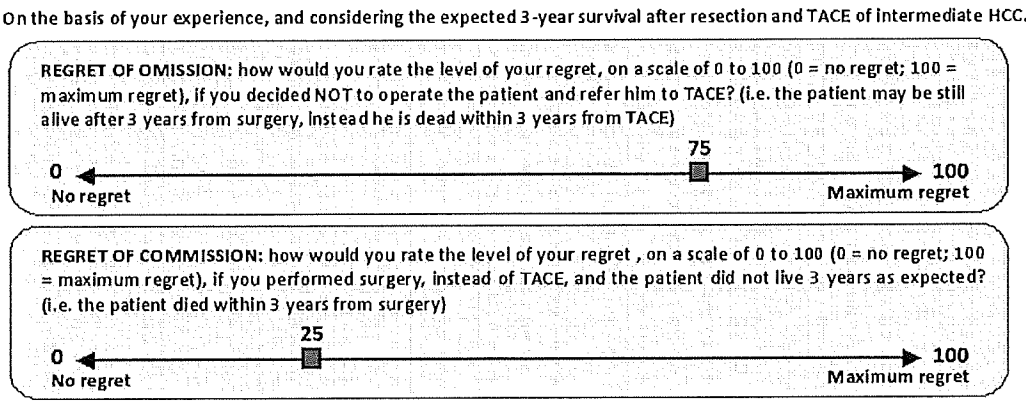


Fig. 1. The DVAS was used to obtain the decision maker's P_t as reported in Equation (1). The questions depicted were associated with each of the three clinical cases of the survey conducted in the present study.

Address reprint requests to: Alessandro Cucchetti, M.D., Policlinico Sant'Orsola-Malpighi, University of Bologna, Via Massarenti 9, 40138 Bologna, Italy. E-mail: aleqko@libero.it; fax: +39-051-304902.

View this article online at wileyonlinelibrary.com.

DOI 10.1002/hep.27321

Potential conflict of interest: Dr. Cillo consults and advises Novartis and advises Astellas. Dr. Piscaglia advises, is on the speakers' bureau of, and received grants from Bayer.

equal to or above this threshold probability, because surgical choice would be associated with the least amount of regret. This can be summarized as shown by Equation 1[13,17]:

$$P_t = 1/(1+(\text{regret of omission}/\text{regret of commission})) \quad (1)$$

where P_t = threshold probability. Regret of omission was defined as regret felt by the physician (surgeon or hepatologists) who withheld hepatectomy from a patient with intermediate HCC who, otherwise, may have benefited from this treatment. Regret of commission was defined as regret felt by the physician who decided to perform a hepatectomy, instead of referring the patient to TACE (Fig. 1). For example, if a physician feels that his or her regret of omission is 4 times greater than regret of commission, the threshold probability at which he or she is indifferent between liver resection and TACE is equal to 20%. This means that he or she should offer hepatectomy if the probability of survival after surgery, at a predefined temporal endpoint, is greater than 20%; otherwise, surgery should be avoided and TACE would represent the least regretful strategy.

The clinical question regarding which treatment to adopt can be broken into three categories: (1) a physician can be aggressive and recommend resection to all patients with intermediate HCCs; (2) a physician can decide to withhold surgery and administer TACE to all patients; or (3) a physician can use a prediction model for guidance. The decision tree that describes the present clinical problem is depicted in Fig. 2. In the present study, we employed a Regret-DCA model to compute the expected regret for a range of threshold probabilities. Regret-DCA is a modification of the original DCA, based on expected regret. Assuming that the threshold probability of an event at which a physician would opt for treatment is informative of how the physician weighs the relative harms of a false-positive (FP) and a false-negative (FN) prediction, the expected regret of the model, across different threshold probabilities, can be derived and computed as follows[13-15,18] (Equations 2-4):

$$\text{Expected regret} = (FP/n)\times(P_t/1-P_t)+(FN/n) \quad (2)$$

$$\text{Expected regret[Surgery]} = (1-s)\times(P_t/1-P_t) \quad (3)$$

$$\text{Expected regret [No Surgery]} = s \quad (4)$$

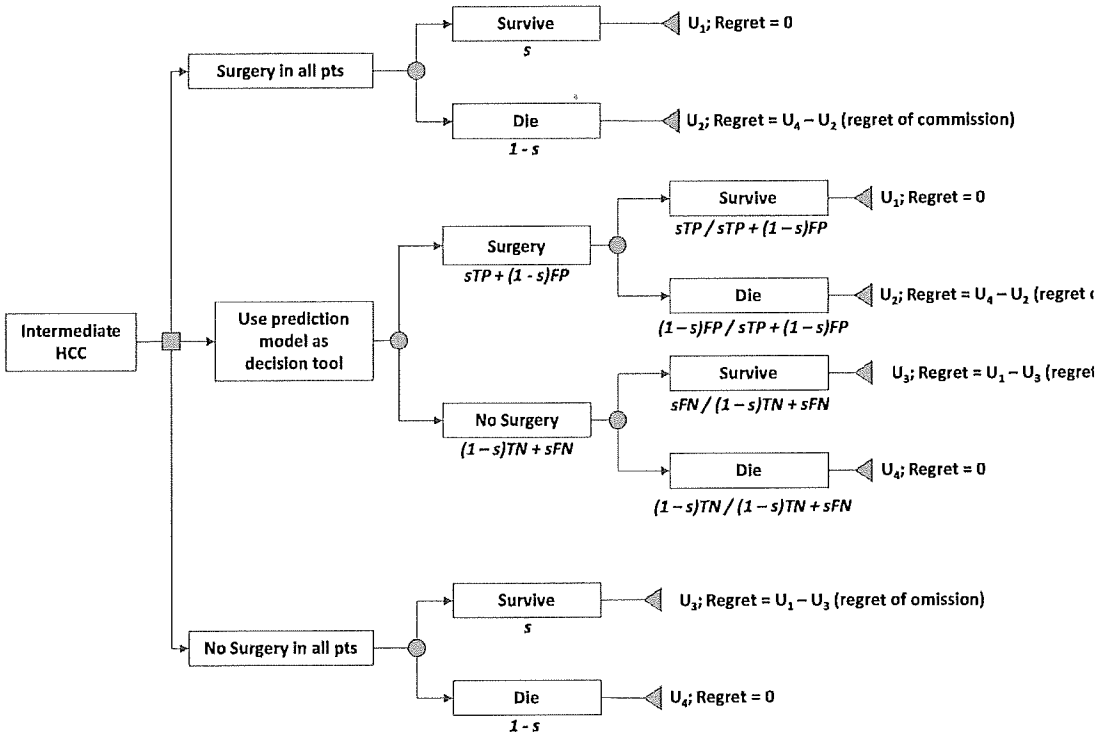


Fig. 2. A decision tree for performing hepatic resection in patients with intermediate HCC, where s = 3-year survival expected after surgery in the whole study population, TP = true positive, TN = true negative, FP = false positive, and FN = false negative. U_i are the utilities associated with each outcome. Mathematical details of the regret-based decision tree were previously published.[13-15]

where P_t = threshold probability (from Equation (1)), FP = false positive, that is, the number of patients who will die within a specific temporal endpoint and for whom the estimated probability of survival is equal or greater than the threshold probability, FN = false negative, that is, the number of patients who will survive longer than a specific temporal endpoint and for whom the estimated probability of survival was lower than the threshold probability, n = number of patient in the DCA, and s = 3-year survival estimation. In the present analysis, the temporal endpoint was set at 3 years, which corresponds to the median survival of the patient study population. Once Regret-DCA is plotted, the physicians can consult a predictive model to assess the probabilities of survival based on specific clinical and tumor features. For this reason, a Cox's regression model was developed on 247 patients with cirrhosis referred to surgery for intermediate HCC, as detailed in the next paragraph.

Study Population. Between January 1999 and November 2011, 806 consecutive patients with cirrhosis were resected for HCC at two tertiary referral surgical units; the policies of the two centers regarding indications for HR have already been published.[19,20] From the two institutional prospective databases, 247 patients with cirrhosis, resected for intermediate HCC, were selected for the present analysis. Intermediate HCC was defined on the basis of BCLC classification as follows: single tumor more than 5 cm in diameter; 2-3 tumors, of which at least 1 was >3 cm in diameter; more than 3 tumors of any diameter; and absence of extrahepatic metastasis and absence of tumor invasion into a major branch of the portal or hepatic veins and performance status 0.[5-8] This definition is supported by the decision-making nature of the present analysis (that means that decision for resection or TACE is made on such criteria) and by recent surgical literature.[5-8]

Other exclusion criteria were the following: patients treated as an emergency; those receiving portosystemic shunts before (or at the same time as) HR or those undergoing preoperative portal vein embolization; evidence of direct invasion of adjacent organs; or spread to the lymph nodes of the hepatic hilum at pathological examination. Presence of a positive margin (R^+) at pathological examination represented additional exclusion criteria and all resections considered here were curative resections (R0). The diagnosis of cirrhosis was confirmed on histological specimens. All patients underwent intraoperative hepatic ultrasonography and were deemed to have resectable tumors at the time of surgery. Demographic, clinical, and tumoral characteristics of the surgical patients are reported in Table 1. Radiological tumoral characteristics are considered, not those observed on the pathological specimen.

Table 1. Baseline Characteristics of Patients With Cirrhosis Undergoing Hepatic Resection for Intermediate HCC

Variable	In Study (n = 247)
Age, years	65 (57-71)
Male gender (%)	201 (81.4)
$HBsAg^+$ (%)	57 (23.1)
$Anti\text{-}HCV^+$ (%)	126 (51.0)
Mild ascites (%)	24 (9.7)
Presence of varices (%)	58 (23.5)
Serum albumin, g/dL	3.8 (3.4-4.0)
Total bilirubin, mg/dL	0.85 (0.59-1.25)
Platelet count, $\times 10^3$/mmc	149 (105-218)
INR	1.13 (1.07-1.21)
Child-Pugh score	5 (5-6)
A5 (%)	141 (57.1)
A6 (%)	86 (34.8)
B7 (%)	18 (7.3)
B8 (%)	2 (0.8)
MELD score	8 (7-9)
Radiological tumor number	1 (1-2)
Single tumor (%)	124 (50.2)
Two or three tumors (%)	93 (37.7)
More than three tumors (%)	30 (12.1)
Radiological largest tumor size, cm	6.0 (5.0-7.7)
OS	
1 year (95% CI)	77.8% (72.1-82.6)
3 year (95% CI)	48.7% (41.4-55.5)
5 year (95% CI)	33.8% (26.2-41.5)

Continuous variables are reported as medians and IQRs (25th-75th percentiles).

Abbreviations: HBsAg, hepatitis B surface antigen; HCV, hepatitis C virus.

Statistical Analysis. Clinical and tumoral characteristics were reported as the number of cases and prevalence, and median and interquartile ranges (IQRs), as appropriate, on the basis of type of variables and distributions. Median and IQR were selected, instead of mean and standard deviation, because of violation of normal distribution by the majority of variables (by Kolmogorov-Smirnov's test). Comparisons were performed using appropriate statistical analyses on the basis of type of variable or distribution. Overall survival (OS) was estimated using Kaplan-Meier's method. Each variable was entered in a univariate Cox's regression model. To avoid any colinearity within the model, albumin and bilirubin were not included in the multivariate model because they were already incorporated in the Child-Pugh classification; in addition, because the international guidelines are mainly based on Child-Pugh classification for the assessment of liver functional reserve, the Model for End-stage Liver Disease (MELD) score was also removed from the multivariate analysis and here reported only as descriptive variable of the study

population. Otherwise, any vatiables having a P value <0.10 based on univariate Cox's regression were entered in the multivariate backward proportional hazard model. The baseline cumulative hazard after surgery was assessed by the means of Breslow's estimator. Schoenfeld's residuals confirmed the proportional hazard assumption of the model. Once beta-coefficients (b) for each independent ptedictor (x) and baseline cumulative hazard values were obtained from multivariate regression, the 3-year survival estimations was calculated, and reported, using the following equation: S0(year)^(EXP(PI)), where S0(year) = -EXP(baseline cumulative hazard at 3 years) and $PI=x_kb_1+x_2b_2+\ldots..+x_kb_k$. The 3-year survival probabilities predicted from the model, for each of the 247 patients, were used to compute the Regret-DCA. Statistical analyses were performed with the STATA statistical program (StataCorp LP, College Station, TX) and R (version 2.12.0; R Foundation for Statistical Computing, Vienna, Austria).

Description of the Survey. Regrets of omission and commission were assessed using an online survey. This assessment was obtained using the dual visual analog scale (DVAS), as repotted in Fig. 1. The DVAS comprise of two 100-point scales, each anchored to the absence of regret and maximum regret.[13-15] One of the scales was used to measure the regret of omission and the other to measure the regret of commission. Three hypothetical clinical cases were sent for consultation to physicians involved in HCC management (surgeons or hepatologists) with at least 10 years of experience in the treatment of HCC and cirrhosis. These clinical cases wete based on the expert opinion with awareness of a recent proposal for a subclassification of intermediate HCCs.[21] The base case was tepresented by a 70-year-old man with hepatitis C cirrhosis and HCC beyond the early BCLC stage, without vascular invasion or distant metastases, surgically removable, having an Eastern Cooperative Oncology Group performance status of 0.

Case 1: Laboratory shows: total bilirubin = 0.9 mg/dL; albumin = 3.8 g/dL; international normalized ratio (INR) = 1.3; no ascites; no encephalopathy; and Child-Pugh A5. A computed tomography (CT) scan shows two nodules: The latgest is 40 mm and the smallest is 10 mm, both located in livet segment 7. There are no varices at endoscopy, and platelet count is normal.

Case 2: Laboratory shows: total bilirubin = 1.8 mg/dL; albumin = 3.0 g/dL; INR = 1.4; no ascites; no encephalopathy; and Child-Pugh A6. There are varices at endoscopy, there is splenomegaly at ultrasound (US) examination, and the platelet count is between 100 and 120 $\times$ 10^3/mm^3. A CT scan shows two nodules: The largest is 60 mm and the smallest is 35 mm, located in different liver segments of the same hemiliver.

Case 3: Laboratory shows: total bilirubin = 1.9 mg/dL; albumin = 2.6 g/dL; INR = 1.6; no ascites; no encephalopathy; and Child-Pugh B7. There are varices at endoscopy, splenomegaly at US, and the platelet count is approximately 90 $\times$ 10^3/mm^3. At CT scan, there is one nodule of 70 mm.

Physicians involved were asked to assess their personal regret of omission and commission in terms of 3-year survival after resection, and considering, as an alternative strategy to withhold surgery, TACE. Threshold probabilities were compared with the Regret-DCA and with the predicted survivals obtained from the model in order to assess whether each single physician would recommend surgery or TACE in relationship with his or her specific regret and estimated probability of survival. We did not include the specific survival data in any of the vignette description, but assumed that it would be obvious to the experienced physicians that survival differed among all 3 cases described in the survey. Physicians were not asked for their own decision because we were mostly interested whether the physicians' preferences would cluster within three zones defined by the Regret-DCA: perform surgery on all patients; avoid surgery on all patients and administer TACE; and make a decision based on the prognostic model.

Results

Patient Characteristics and Outcome. Baseline characteristics of the 247 patients with cirrhosis resected for intermediate HCC are reported in Table 1. During a median follow-up of 24 months (range, 1 month to 11 years), 132 patients died (53.4%). The median survival was 35 months (95% confidence interval [CI]: 26-42); the 1-, 3-, and 5-year survival rates were 77.8%, 48.7%, and 33.8%, respectively. Results from Cox's regression are reported in Table 2. Child-Pugh score ($P = 0.001$), number of tumors (single vs. two or three vs. multiple nodules; $P = 0.015$) and presence of esophageal varices (EV) at endoscopy ($P = 0.015$) were found to be independent predictors of OS after hepatic resection. These variables were used to build the 3-year probability of survival for each of the 247 patients.

Regret-DCA. Regret-DCA was applied to the 3-year survival prediction, and Fig. 3 depicts how the

Table 2. Uni- and Multivariate Cox's Regressions on OS of Patients Submitted to Hepatic Resection for Intermediate HCC

Variable	Univariate Regression		Multivariate Regression	
	Exp (B) (95% CI)	*P* Value	Exp (B) (95% CI)	*P* Value
Age (per year increase)	1.006 (0.987-1.025)	0.547	–	–
Gender (male vs. female)	0.815 (0.529-1.257)	0.355	–	–
HBsAg (positive vs. negative)	1.154 (0.778-1.711)	0.478	–	–
Anti-HCV (positive vs. negative)	1.319 (0.926-1.879)	0.125	–	–
Mild ascites (present vs. absent)	1.200 (0.661-2.178)	0.549	–	–
EV (present vs. absent)	1.667 (1.138-2.443)	0.009	1.535 (1.044-2.257)	0.029
Serum albumin (per g/dL increase)	0.551 (0.388-0.784)	0.001	–	–
Total bilirubin (per mg/dL increase)	1.959 (1.407-2.729)	0.001	–	–
Platelet count (per $\times 10^3$/mmc increase)	0.998 (0.996-1.001)	0.069	0.999 (0.997-1.002)	0.583
INR (per unit increase)	2.568 (0.662-9.958)	0.173	–	–
Child-Pugh score (per unit increase)	1.592 (1.216-2.085)	0.001	1.560 (1.187-2.052)	0.001
MELD score (per unit increase)	1.206 (1.088-1.336)	0.001	–	–
Radiological tumor number (single vs. 2-3 vs. >3)	1.345 (1.053-1.718)	0.018	1.357 (1.061-1.737)	0.015
Radiological largest tumor size (per cm increase)	1.008 (0.946-1.073)	0.815	–	–

To avoid any colinearity within the model, albumin and bilirubin were not included in the multivariate model because they were already incorporated in the Child-Pugh classification; in addition, because international guidelines are mainly based on Child-Pugh classification for the liver functional reserve assessment, MELD score was also removed from the multivariate analysis. Otherwise, any variables having a *P* value <0.10 at univariate Cox's regression were entered in the multivariate backward proportional hazard model. The 3-year baseline cumulative hazard was 0.034.

Abbreviations: HBsAg, hepatitis B surface antigen; HCV, hepatitis C virus.

expected regret changes for different P_t values. As shown, for P_t <70%, the least regretful strategy was to employ the model shown in Table 3 and act accordingly. If, for a given threshold, the model's prediction of the probability of survival is greater than the P_t value, then the physician should offer hepatectomy. If the estimated probability of survival is smaller than the threshold probability, then withholding surgery and administering TACE is the least regretful strategy. For the P_t above 70%, the most optimal, least-regretful strategy is to withhold surgery and offer TACE to all patients. For P_t values below 35%, an alternative should be to offer surgery to all patients, regardless of the model's prediction, because regret of making the

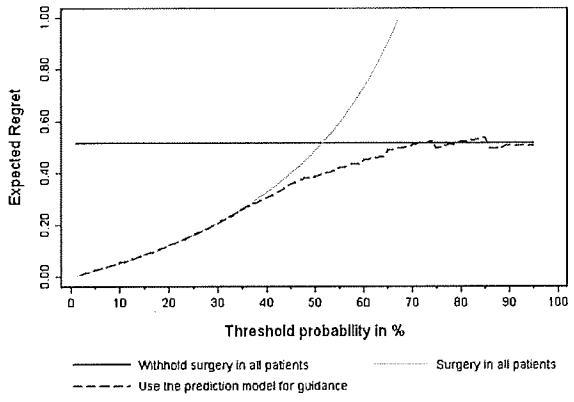


Fig. 3. Regret-DCA for the Cox's model built to predict the 3-year survival probability based on records of 247 patients resected for intermediate HCC. The optimal strategy is the action that results in the least amount of regret in case it is proven wrong. For P_t <70%, the least regretful strategy is to consult the prediction model for treatment selection. According to the prediction model, a patient should have surgery if the survival probability is greater than the threshold probability. The patient should not be referred to surgery and should be recommended TACE otherwise. For P_t values <35%, the prediction model remains the best decision-making approach because it is equal in regret to the strategy to submit all patients to surgery.

Table 3. Results From Regression Analysis of 247 Patients Resected for Intermediate HCC

Clinical Scenario	OS		
	1-year (%)	3-year (%)	5-year (%)
Child-Pugh A5, no varices			
Single large nodule	86.9	65.3	51.7
Two or three nodules	82.7	56.1	40.8
More than three nodules	77.2	45.6	29.6
Child-Pugh A5, with varices			
Single large nodule	80.6	52.0	36.3
Two or three nodules	74.7	41.1	25.2
More than three nodules	67.3	29.9	15.4
Child-Pugh A6, no varices			
Single large nodule	80.4	51.4	35.7
Two or three nodules	74.3	40.6	24.7
More than three nodules	66.8	29.4	15.0
Child-Pugh A6, with varices			
Single large nodule	71.5	36.0	20.5
Two or three nodules	63.4	25.0	11.7
More than three nodules	53.8	15.2	5.4
Child-Pugh B7, no varices			
Single large nodule	71.1	35.4	20.0
Two or three nodules	62.9	24.5	11.3
More than three nodules	53.3	14.8	5.2
Child-Pugh B7, with varices			
Single large nodule	59.2	20.3	8.5
Two or three nodules	49.1	11.5	3.5
More than three nodules	38.1	5.3	1.1

Regret-DCA was performed on the 3-year overall survival (OS) prediction. The 1- and 5-year survival predictions are reported here for descriptive purposes.

choice based on the model versus "surgery to all" is equal. An example can help clarify the process. Assume that the P_t value calculated by a physician for the 3-year survival for a specific HCC patient is 46.2% (regret of omission = 70; regret of commission = 60) and assume that the disease is a Child-Pugh class A5, without varices, and with a single large HCC: in this case, the decision process requires comparison of the physician's threshold to the prediction probability obtained from the regression model described in Table 3. As can be noted, the predicted 3-year survival is 65.3%, thus, higher than the physician threshold of 46.2%. Consequently, surgery represents the least regretful treatment strategy.

Survey Results. To further explain this decision approach and assess a distribution of the threshold probabilities in individual physicians, a survey was conducted among 40 physicians, asking for regret assessment of the three clinical cases previously described. Detailed results are reported in Table 4, and a summary is reported in Table 5. There were 120 decisions: HR would have been recommended in 41 (34%), whereas TACE would have been recommended in 79 (66%) cases. As can be noted from Table 4, threshold probabilities (P_t) for case 1 ranged considerably, from 5% to 58.8%; the estimated probability of survival in this case was 56.1%. Basing the final decision on the regret model, 95% of physicians (38 of

Table 4. Results of the Survey Among 40 Physicians With Expertise in HCC Treatment

		P_t					
Physician	Specialty	Case 1 (%)	Predicted Decision	Case 2 (%)	Predicted Decision	Case 3 (%)	Predicted Decision
1	Surg	11.1	HR	30.0	TACE	33.3	TACE
2	Hep	40.0	HR	73.3	TACE	87.5	TACE
3	Sur	9.1	HR	16.7	HR	50.0	TACE
4	Hep	40.0	HR	70.0	TACE	75.0	TACE
5	Sur	5.0	HR	76.2	TACE	90.0	TACE
6	Sur	18.2	HR	36.4	TACE	40.0	TACE
7	Hep	10.0	HR	50.0	TACE	80.0	TACE
8	Sur	5.0	HR	44.4	TACE	60.0	TACE
9	Sur	31.0	HR	33.3	TACE	50.0	TACE
10	Hep	22.2	HR	85.7	TACE	94.1	TACE
11	Hep	42.9	HR	63.6	TACE	54.5	TACE
12	Hep	28.6	HR	90.0	TACE	95.2	TACE
13	Sur	33.3	HR	89.5	TACE	90.9	TACE
14	Hep	9.1	HR	66.7	TACE	90.5	TACE
15	Hep	20.0	HR	75.0	TACE	55.6	TACE
16	Sur	25.0	HR	41.7	TACE	40.0	TACE
17	Hep	37.5	HR	57.1	TACE	71.4	TACE
18	Sur	11.1	HR	60.0	TACE	80.0	TACE
19	Sur	18.2	HR	20.0	HR	62.5	TACE
20	Sur	20.0	HR	81.8	TACE	86.4	TACE
21	Sur	15.8	HR	60.0	TACE	85.0	TACE
22	Hep	30.8	HR	66.7	TACE	70.0	TACE
23	Hep	12.5	HR	40.0	TACE	70.0	TACE
24	Sur	22.2	HR	33.3	TACE	31.8	TACE
25	Sur	40.0	HR	47.4	TACE	33.3	TACE
26	Hep	19.0	HR	89.5	TACE	95.2	TACE
27	Sur	20.8	HR	35.0	TACE	50.0	TACE
28	Hep	20.0	HR	50.0	TACE	70.0	TACE
29	Sur	20.0	HR	57.1	TACE	87.5	TACE
30	Hep	30.0	HR	80.0	TACE	50.0	TACE
31	Hep	58.3	TACE	78.9	TACE	78.9	TACE
32	Hep	58.8	TACE	77.8	TACE	88.9	TACE
33	Hep	18.2	HR	33.3	TACE	82.4	TACE
34	Hep	20.0	HR	30.0	TACE	70.0	TACE
35	Sur	15.8	HR	38.9	TACE	68.2	TACE
36	Hep	40.0	HR	60.0	TACE	90.0	TACE
37	Hep	10.0	HR	30.0	TACE	69.2	TACE
38	Sur	16.7	HR	20.0	HR	50.0	TACE
39	Sur	46.7	HR	53.8	TACE	83.3	TACE
40	Sur	13.6	HR	75.0	TACE	80.0	TACE

If the predicted survival is greater than the P_t, the optimal choice would be to perform HR, otherwise TACE should be administered. Predicted survivals at 3 years: case 1, 56.1%; case 2, 25.0%; case 3, 20.3%.

Abbreviations: Sur, surgeon; Hep, hepatologist.

Table 5. Summary Statistics of the Survey Conducted Among Surgeons and Hepatologists to Test the Regret Theory

Clinical Scenario	All Physicians (n = 40)	Surgeons (n = 20)	Hepatologists (n = 20)	*P* Value
Case 1				
P_t	20.0 (14.2-32.8)	18.2 (11.7-24.3)	25.4 (18.4-40.0)	0.076
Regret of omission	80 (70-90)	90 (80-95)	75 (60-90)	0.006
Regret of commission	20 (15-40)	20 (10-30)	20 (15-40)	0.277
Case 2				
P_t	57.1 (35.3-75.0)	43.1 (33.3-60.0)	66.7 (50.0-78.7)	0.024
Regret of omission	40 (20-60)	50 (30-70)	30 (20-50)	0.049
Regret of commission	50 (30-70)	40 (25-60)	60 (40-70)	0.052
Case 3				
P_t	70.7 (51.1-87.2)	61.3 (42.5-84.6)	77.0 (70.0-89.7)	0.028
Regret of omission	25 (10-40)	30 (10-60)	20 (10-30)	0.127
Regret of commission	70 (50-80)	50 (30-79)	75 (70-90)	0.026

Values are reported as medians and IQRs (25th-75th percentiles).

40) would theoretically decide for the same treatment, namely, HR. In 31 cases, the preferences fell within a zone "perform surgery in all patients" on the Regret-DCA curve (Fig. 3). Regarding case 2, threshold probabilities ranged from 16.7% to 90.0% and the use of the Regret-DCA would have led 92.5% of physicians to decide for TACE (estimated survival probability: 25.0%). Regarding case 3, threshold probabilities ranged between 31.8% and 95.2%, possibly leading all physicians to decide for TACE (estimated survival probability: 20.3%). Table 5 reports a summary of the survey. It was of interest to observe that median threshold probabilities of surgeons were always lower than that of hepatologists. Surgeons reported a statistically significant higher regret of omission than hepatologists for cases 1 ($P = 0.006$) and 2 ($P = 0.049$), whereas the magnitude of this difference for case 3 was less pronounced ($P = 0.127$). Conversely, hepatologists suffer from higher regret of commission, especially for case 3 ($P = 0.026$), whereas for case 2, the difference, in comparison to surgeons, was smaller ($P = 0.052$), or even null for case 1 ($P = 0.277$).

Discussion

Making decisions, which are almost always done under uncertainty, represents one of the essential aspects of clinical medicine. Theoretically, no decision can be guaranteed to be absolutely correct.[22] As a result, decisional tasks may be subject to errors that can have important consequences regarding patient outcomes. After making a decision, some can have the inclination to focus on the negative aspects of the choice made[23]; thus, they can experience regret. The cognitive error that leads to regret is fundamentally influenced by uncertainty about best decisions and expected outcomes. In the field of HCC, there is still a substantial uncertainty regarding optimal diagnostic processes and therapeutic approach, especially in the setting of intermediate stage.[5,8,21,24] This is mainly the consequence of the absence of high-level evidence regarding the superiority of one treatment over the alternative.[21,24] Therefore, the decision-making approach to HCC is well suited for application of regret theory methodology.

The present results suggest that regret theory can be applied to decision making of surgery versus avoid surgery, and administer TACE to patients with intermediate HCC. Both European and American guidelines on the management of HCC currently recommend TACE as the first-line therapy for this tumor stage.[2,3] However, patients in this tumor stage are highly heterogeneous, and it seems unlikely that one single treatment can best serve such a mixed patient group.[21] In addition, given the high stakes, this is ultimately a patient-driven, preference-based decision. However, in practice, the hepatologist's and surgeon's preferences will determine the approach.[5-7,25] The present Regret-DCA shows that the optimal therapeutic option can effectively include both HR and TACE depending on the decision maker's preferences expressed here by the regret threshold values. Neither surgery nor TACE are *a priori* excluded in the treatment strategy of intermediate HCC, but are related to the feeling of each single physician involved in the decision making, based on his or her clinical skills and scientific evidence knowledge. Modern cognitive theories suggest that the decision-making process includes both so-called type 1 (based on feeling, intuition, and automatic responses) and type 2 processes (which include logical analysis and deliberation).[26] Regret is cognitive emotion and, as such, may include both the analytical assessment of the expected outcomes, but also explicitly pushes physicians to evaluate and anticipate their personal feeling in the case of a wrong decision.[27,28] Thus,

both type 1 and 2 cognitive components are here taken into account in a comprehensive approach to the decision making. By weighting the personal regret about surgery, in comparison to the alternative treatment, surgeons and/or hepatologists can reach the final therapeutic decision on the basis of expected outcome and correlation with the least amount of regret. Consequently, the derived choice would be considered optimal from both analytical and emotional points of view. We believe that the regret-based approach is probably applicable to clinical situations with high stakes, such as the one considered here; other types of decisions are probably better addressed by the more-complex dual-processing model, in which regret is only one, albeit important, of the features of decision making.[11,13-15]

The results from the survey reported here confirm the utility of the Regret-DCA. The regret threshold probabilities were found to vary across a large range of values: This aspect highlights the fact that different physicians can prescribe different therapies. Nevertheless, the application of the regret DCA shows that virtually all physicians can act according to the regret-threshold model leading to similar decisions regarding which treatment to adopt, even in the presence of high uncertainty. This finding is of particular interest. In fact, it can be expected that surgeons and hepatologists could have different threshold probabilities, and, from the present study, it is clear that surgeons were more prone to experience regret of omission, in comparison to hepatologists. This is a feature fundamentally intrinsic to being a surgeon, and this difference emphasizes the need for a multidisciplinary approach in the decision making of HCC, given that the surgeon's regret may fundamentally differ from the hepatologists' (and the patient's) regret. Our results are of interest to the guideline developers, particularly in the setting of the development of preference-based recommendations. The guidelines panels typically develop their recommendations for "an average" patient with HCC based on the literature data. However, the actual clinical characteristics of the patients observed in the clinics often differ from the literature in important ways, and, most important, their preferences for the treatments can often significantly differ from the preferences of the members of the guidelines panels. Our analysis indicates that it is possible to identify some zones of preferences where recommendations will likely be uniformly accepted versus those where individual preferences of both surgeon and patient will vary from case to case. This approach can better individualize treatment choices, making them tailored on the peculiarities of each patient's characteristics and needs.

The present study has limitations that deserve discussion. First, we were forced to develop our own surgical prognostic model because no similar tools are available in the literature.[5-7] Further larger and externally validated prognostic models for the intermediate stage of HCC are warranted, but the one we developed fit the purpose of the present analysis. Second, the issue of treatment of HCC is more complex than summarized in the present model. Many factors, such as the availability of liver transplantation after TACE or resection, the ability of surgeons and radiologists in providing the two alternative strategies as well as other clinical aspects, play a role in the feeling of regret that a physician could experience. Moreover, no direct comparison between surgery and TACE was available for further investigation of the regret theory. However, in the final analysis, it emerged that most of these factors converge to the question of what is the optimal choice under clinical uncertainty. Therefore, the application of the regret decision theory, even at risk of simplifying the real-world situations, appears to be well suited to analyze the problem addressed in the present analysis. It has also to be noted that data regarding TACE are not needed for the present analysis, because this procedure remained the comparison in the survey where physicians were asked to elicit their regret in relationship with resection to TACE. Third, we have not actually asked the physicians who participated in the survey to indicate their real choices regarding surgery versus TACE; our assessments of their choices were based on the theoretical predictions based on their response to questions shown in Fig. 1. Whether the physicians make their actual decisions according to the regret threshold model remains to be demonstrated, but this aspect is worthy of a specific and dedicated study. The DCA model assumes that preferences (as elicited by regret in our case) and the predicted probabilities of survival are independent. Given the wide range of the threshold probabilities observed among the physicians we surveyed, it does not seem likely that regret values and survival probabilities correlate. Therefore, change in the median values of regret observed among the three vignettes reflects the physicians' preferences, rather than estimated probability of survival.

In conclusion, we provided here evidence that the regret theory can be applied in the setting of treatment of intermediate HCC, helping physicians to minimize different points of view in favor of the best treatment choice for each single patient. Our results also support

the recommendation for a multidisciplinary approach, which can minimize the amount of regret about the therapeutic choice made. The uncertainty highlighted by this study, and by the survey conducted, suggests the urgent need for high-level evidence of the superiority of one treatment versus the alternative in the setting of intermediate HCC. However, obtaining better evidence will not per se lead to better decision making. We believe that wider application of the methods described in this article may actually improve physicians' decision making and, in turn, patients' outcomes.

References

1. El-Serag HB. Epidemiology of viral hepatitis and hepatocellular carcinoma. Gastroenterology 2012;142:1264-1273.
2. European Association for the Study of the Liver; European Organisation for Research and Treatment of Cancer. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol 2012;56:908-943.
3. Bruix J, Sherman M; American Association for the Study of Liver Diseases. Management of hepatocellular carcinoma: an update. HEPATOLOGY 2011;53:1020-1022.
4. Forner A, Reig ME, de Lope CR, Bruix J. Current strategy for staging and treatment: the BCLC update and future prospects. Semin Liver Dis 2010;30:61-74.
5. Torzilli G, Belghiti J, Kokudo N, Takayama T, Capussotti L, Nuzzo G, et al. Snapshot of the effective indications and results of surgery for hepatocellular carcinoma in tertiary referral centers: is it adherent to the EASL/AASLD recommendations?: an observational study of the HCC East-West study group. Ann Surg 2013;257:929-937.
6. Zhong JH, Ke Y, Gong WF, Xiang BD, Ma L, Ye XP, et al. Hepatic resection associated with good survival for selected patients with intermediate and advanced-stage hepatocellular carcinoma. Ann Surg 2013 Oct 3. [Epub ahead of print] PMID: 24096763.
7. Wang JH, Changchien CS, Hu TH, Lee CM, Kee KM, Lin CY, et al. The efficacy of treatment schedules according to Barcelona Clinic Liver Cancer staging for hepatocellular carcinoma—survival analysis of 3892 patients. Eur J Cancer. 2008;44:1000-1006.
8. Torzilli G, Belghiti J, Kokudo N, Takayama T, Capussotti L, Nuzzo G, et al. Reply to the letter to the editor entitled "A snapshot of the effective indications and results of surgery for hepatocellular carcinoma in tertiary referral centers: is it adherent to the EASL/AASLD recommendations? An observational study of the HCC East-West Study Group": when the study setting "ignores" the patients. Ann Surg 2013 Nov 18. [Epub ahead of print] PMID: 24253157.
9. Loomes G, Sugden R. Regret theory: an alternative theory of rational choice under uncertainty. Econ J 1982;92:805-824.
10. Smith RD. Is regret theory an alternative basis for estimating the value of healthcare interventions? Health Pol 1996;37:105-115.
11. Djulbegovic B, Hozo I, Schwartz A, McMasters KM. Acceptable regret in medical decision making. Med Hypotheses 1999;53:253-259.
12. Gilovich T, Medvec VH. The experience of regret: what, when, and why. Psychol Rev 1995;102:379-395.
13. Tsalatsanis A, Hozo I, Vickers A, Djulbegovic B. A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making. BMC Med Inform Decis Mak 2010;10:51.
14. Tsalatsanis A, Barnes LE, Hozo I, Djulbegovic B. Extensions to regret-based decision curve analysis: an application to hospice referral for terminal patients. BMC Med Inform Decis Mak 2011;11:7.
15. Hernandez JM, Tsalatsanis A, Humphries LA, Miladinovic B, Djulbegovic B, Velanovich V. Defining optimum treatment of patients with pancreatic adenocarcinoma using regret-based decision curve analysis. Ann Surg 2013 Oct 28. [Epub ahead of print] PMID: 24169177.
16. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. N Engl J Med 1975;293:229-234.
17. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980;302:1109-1117.
18. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565-574.
19. Cillo U, Vitale A, Grigoletto F, Farinati F, Brolese A, Zanus G, et al. Prospective validation of the Barcelona Clinic Liver Cancer staging system. J Hepatol 2006;44:723-731.
20. Cucchetti A, Piscaglia F, Cescon M, Ercolani G, Terzi E, Bolondi L, et al. Conditional survival after hepatic resection for hepatocellular carcinoma in cirrhotic patients. Clin Cancer Res 2012;18:4397-4405.
21. Bolondi L, Burroughs A, Dufour JF, Galle PR, Mazzaferro V, Piscaglia F, et al. Heterogeneity of patients with intermediate (BCLC B) Hepatocellular Carcinoma: proposal for a subclassification to facilitate treatment decisions. Semin Liver Dis 2012;32:348-359.
22. Djulbegovic B, Hozo I. When should potentially false research findings be considered acceptable? PLoS Med 2007;4:e26.
23. Aronson E, Wilson TD, Akerr RM. Social Psycology, 3rd ed. New York: Longman; 1999.
24. Italian Association for the Study of the Liver (AISF); AISF Expert Panel; AISF Coordinating Committee. Position paper of the Italian Association for the Study of the Liver (AISF): the mulùdisciplinary clinical approach to hepatocellular carcinoma. Dig Liver Dis 2013;45:712-723.
25. Hsu CY, Hsia CY, Huang YH, Su CW, Lin HC, Pai JT, et al. Comparison of surgical resection and transarterial chemoembolization for hepatocellular carcinoma beyond the Milan criteria: a propensity score analysis. Ann Surg Oncol 2012;19:842-849.
26. Kahneman D. Maps of bounded rationality: psychology for behavioral economics. Am Econ Rev 2003;93:1449-1475.
27. Zeelenberg M, Pieters R. A theory of regret regulation 1.1. J Consumer Psychol 2007;17:29-35.
28. Slovic P, Finucane ML, Peters E, MacGregor DG. Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. Risk Anal 2004;24:311-322.

Gil-Herrera *et al. BMC Medical Informatics and Decision Making* (2015) 15:98
DOI 10.1186/s12911-015-0216-9

BMC
Medical Informatics & Decision Making

RESEARCH ARTICLE Open Access

Rough set theory based prognostic classification models for hospice referral

Eleazar Gil-Herrera[1†], Garrick Aden-Buie[2†], Ali Yalcin[2*], Athanasios Tsalatsanis[1], Laura E. Barnes[3] and Benjamin Djulbegovic[1,4]

Abstract

Background: This paper explores and evaluates the application of classical and dominance-based rough set theory (RST) for the development of data-driven prognostic classification models for hospice referral. In this work, rough set based models are compared with other data-driven methods with respect to two factors related to clinical credibility: accuracy and accessibility. Accessibility refers to the ability of the model to provide traceable, interpretable results and use data that is relevant and simple to collect.

Methods: We utilize retrospective data from 9,103 terminally ill patients to demonstrate the design and implementation RST- based models to identify potential hospice candidates. The classical rough set approach (CRSA) provides methods for knowledge acquisition, founded on the relational indiscernibility of objects in a decision table, to describe required conditions for membership in a concept class. On the other hand, the dominance-based rough set approach (DRSA) analyzes information based on the monotonic relationships between condition attributes values and their assignment to the decision class. CRSA decision rules for six-month patient survival classification were induced using the MODLEM algorithm. Dominance-based decision rules were extracted using the VC-DomLEM rule induction algorithm.

Results: The RST-based classifiers are compared with other predictive and rule based decision modeling techniques, namely logistic regression, support vector machines, random forests and C4.5. The RST-based classifiers demonstrate average AUC of 69.74 % with MODLEM and 71.73 % with VC-DomLEM, while the compared methods achieve average AUC of 74.21 % for logistic regression, 73.52 % for support vector machines, 74.59 % for random forests, and 70.88 % for C4.5.

Conclusions: This paper contributes to the growing body of research in RST-based prognostic models. RST and its extensions posses features that enhance the accessibility of clinical decision support models. While the non-rule-based methods—logistic regression, support vector machines and random forests—were found to achieve higher AUC, the performance differential may be outweighed by the benefits of the rule-based methods, particularly in the case of VC-DomLEM. Developing prognostic models for hospice referrals is a challenging problem resulting in substandard performance for all of the evaluated classification methods.

Keywords: Rough set theory, Dominance-based rough set approach, Hospice referral, Prognostic models

*Correspondence: ayalcin@usf.edu
†Equal contributors
[2]Department of Industrial and Management Systems Engineering, University of South Florida, 4202 E. Fowler Avenue, ENB 118, Tampa, FL 33620, USA
Full list of author information is available at the end of the article

Background

Hospice care reduces the emotional burden of illness on terminal patients by optimizing pain relief strategies [1] and provides a demonstrated, cost-effective increase in the quality of end-of-life care when compared to conventional programs [2]. This increase in quality of care elevates the quality of life of both patients and their families [3].

The advantages of hospice care are diminished for terminally ill patients who enter either prematurely or too late. In general, premature hospice referral represents a lost opportunity for the patient to receive potentially effective and life-prolonging treatment. Conversely, late hospice referral is not desirable and negatively impacts both the quality of end-of-life care and the quality of life of patients and their families [4, 5]. According to Medicare regulations, patient eligibility for hospice care is contingent upon a life expectancy of less than six months, as estimated by the attending physician and certified by the medical director of the hospice program [6]. Medicare claims data report that 14.9 % of hospice care patients lived for more than 180 days after enrollment, while 28.5 % were late referrals who died within 14 days [4, 6]. Accurate prognostication of life expectancy is crucial in end-of-life care decisions and is consequently of vital importance for patients, their physicians and their families.

Prognostic models are an important instrument in prognostication as, in conjunction with direct physician observation, they increase the accuracy of prognostication when compared to physician observation alone [7]. However, a significant barrier to the widespread practical use of prognostic models is their perceived lack of clinical credibility [8].

The objective of this work is to explore and evaluate the application of rough set approaches in the development of data-driven prognostic models with respect to two criteria essential to clinical credibility: accuracy and accessibility. To this end, we will explore Rough Set Theory (RST) as it is applied to end-of-life care and hospice referral decision support models. Additionally, we will compare the results of the RST-based models with several widely known methods: logistic regression, support vector machines (SVM), C4.5, and random forests (RF).

This paper is organized as follows: The Motivation Section presents important features of clinically credible prognostic models and other characteristics of clinical data sets that motivate the use of RST. We then present an overview of the fundamental theory of rough sets for analyzing datasets (section Methods), followed with a similar overview of the theory of the Dominance-based Rough Set Approach (DRSA). In this section, we also discuss the use of decision rules in conjunction with the rough set approaches. The section Dataset description describes the dataset used for the demonstration of the proposed prognostic models. Section Experimental design presents the development of the prognostic models, followed by an overview of the performance evaluation methods used in this study. Finally, we report the results and conclusions, and discuss limitations and future directions of our work.

Motivation

The objective of a prognostic model is to determine relationships between covariates and a health-related outcome. In the case of life expectancy estimation, prognostic models improve the accuracy in critical clinical decisions and are shown to be superior to physicians' prognostication alone [9]. Models for estimating the life expectancy of terminally ill patients include the use of statistical and probabilistic methods [10–18], artificial intelligence techniques such as neural networks and support vector machines (SVM) [19–21], decision trees [22, 23] and rough set methods [24, 25]. Survival models [6, 12, 14, 16, 18, 22, 23] focus on estimating the probability that a patient will survive a finite period of time. Classification models, based on methods such as neural networks, SVM and logistic regression [17, 19–21, 26], represent the survival outcome as a binary variable, predicting the status of a patient at a critical point in time (e.g. six months) by classifying the patient as surviving or not surviving the critical time frame. Classification models require the use of non-censored data where survival outcome is known for every patient in the dataset at the critical decision point in time.

A recent review [15] demonstrated that, despite the importance of accurate prognostication within the spectrum of medical care objectives, there is a lack of accessible and accurate prognostic models available to physicians in practice. To withstand clinical trials, and to meet the needs of physicians and patients, a prognostic model must have clinical credibility, meaning that the model must posses a high level of accuracy and accessibility for physicians to believe in the value of the model as a prognostic tool. That is, in addition to accurate prognostication, such a model should be traceable in its structure, meaning the "model's structure should be apparent and its predictions should make sense to the doctors who will rely on them" [8]. Likewise, the model should provide interpretable results that facilitate explanation of the prognosis, the data required for the model must be relevant and simple to collect with high reliability, and physicians must be able to apply the modeling method correctly without violating the fundamental assumptions of the model.

Clinical datasets present unique challenges that must also be addressed when building data-driven prognostic models. Cios and Moore [27] argue that there are a number of features specific to medical data that result from the volume, heterogeneity and complexity of data that

lack canonical form. Additionally, ethical, legal and societal concerns greatly affect the framework under which medical data may be used. The current US model encourages the use of de-identified, minimal risk medical data for research purposes, specifically data collected during routine treatment of patients. It is common for medical data collected in such a way to contain redundant, insignificant, incomplete or inconsistent data objects. Furthermore, the underlying conceptual structures of medicine are not easily formalized mathematically, as the medical field lacks the necessary constraints for the mathematical characterizations common to the physical sciences. As a result, many medical concepts are vaguely defined [28].

Rough Set Theory [29] is a mathematical tool for data analysis that has been used to address vagueness and inconsistencies present in datasets [30]. RST provides a systematic approach for analyzing data without implicit assumptions about relationships between covariates, an advantage that makes RST suitable for integration into medical applications [31]. The information extracted from the dataset by RST and its related methods can be represented in the form of "if–then" decision rules—an intuitive representation that offers significant advantage over "black box" modeling approaches [32] and that increases accessibility and thus clinical credibility.

In the medical field, applications of RST focus mainly on the diagnosis and prognostication of diseases, where it has been demonstrated that RST is useful for extracting medical prognostic rules from minimal information. Tsumoto [33] argues that the concepts of approximation established in RST reflect the characteristics of medical reasoning, explaining why RST performs well in the medical field. For example, RST can be used to highlight non-essential prognostic factors in a particular diagnosis, thus helping to avoid redundant, superfluous or costly tests [34–38]. Recently, methods that combine survival analysis techniques and RST have been used to generate prognostic rules that estimate the survival time of a patient [24, 25].

Methods

Classical rough set approach (CRSA)

Rough Set Theory, introduced by Pawlak in [29], provides methods for generalizing or reducing information so as to facilitate knowledge discovery by exploiting the relational indiscernibility of objects in an information table. Central to RST is the notion that an observed object has a certain amount of information associated with it. When considered in relation to a cohort of observed objects, this information is used to group similar objects into information granules. Together, the information provided by the set of observed objects can be generalized to describe the conditions required for membership in a concept class.

Notation

The methods of classical RST, hereafter referred to as the CRSA, act upon an information table of the form $S = (U, A, V, f)$, where U is a non-empty finite set of objects, called the universe. $A = C \cup \{d\}$ is a set of attributes that describe a given object in U, comprised of a set C of condition attributes and an optional decision attribute d. When d is present, the information table is a decision table. The set of all values, V, contains the value sets V_a, for every attribute $a \in A$. Given an object $x \in U, f: U \times A \rightarrow V$ maps the condition attribute of object x to its associated value $v = f(x, a) \in V_a$. A value attribute pair (a, v) for a given object is referred to as a descriptor.

Table 1 provides an example of a discretized decision table, where six prognostic factors, as the condition attributes, describe seven patients. The decision attribute, presence of coronary disease in the patient, is represented by the binary attribute $d \rightarrow \{\text{Yes}, No\}$.

The objects in a decision table can be grouped according to their descriptors. For example, patients x_5 and x_6 have the same attribute values and are thus indiscernible from each other. In general, two objects $x_i, x_j \in U$ are indiscernible with respect to a set of condition attributes $B \subseteq C$ if $f(x_i, a) = f(x_j, a)\ \forall a \in B$. This relation

Table 1 Example decision table

	Condition attribute[a]						Decision attribute
Patient	c_1 Gender	c_2 Age	c_3 SystBP	c_4 HDL	c_5 Diabetic	c_6 Smoker	d Coronary disease
x_1	F	H	M	L	No	No	No
x_2	M	L	L	L	No	Yes	No
x_3	F	M	M	H	No	No	No
x_4	F	M	M	H	No	No	Yes
x_5	M	H	H	L	Yes	Yes	Yes
x_6	M	H	H	L	Yes	Yes	Yes
x_7	F	M	M	H	No	No	Yes

[a]Gender: Female/Male; Age: L = [54, 59), M = [59, 69), H = [69, 74]; SystBP: L =< 129, M = [129 – 139], H = (139 – 159]; HDL: L = < 40 M = [40 – 60], H => 60

is called an indiscernibility relation, defined as $R(B) = \{(x_i, x_j) \in U : \forall a \in B, f(x_i, a) = f(x_j, a)\}$.

For example, the patients in Table 1 can be separated into four groups according to the indiscernibility relation $R(C)$: $X_1 = \{x_1\}, X_2 = \{x_2\}, X_3 = \{x_3, x_4, x_7\}, X_4 = \{x_5, x_6\}$. These groups of objects are referred to as equivalence classes, or conditional classes for $B \subseteq C$. An equivalence class for the decision attribute is called a decision class or concept, and in this example there are two groups: $Y_{No} = \{x_1, x_2, x_3\}$ and $Y_{Yes} = \{x_4, x_5, x_6, x_7\}$. The equivalence class specified by the object x_i with respect to $R(B)$ is denoted as $[x_i]_B$.

Set approximations

The goal of the CRSA is to provide a definition of a concept according to the values of the attributes of the equivalence classes that contain objects that are known instantiations of the concept. As such, in a consistent decision table, membership in a conditional class implies membership in a particular decision class. In Table 1, $x \in X_4$ implies $x \in Y_{Yes}$. Membership in X_3, however, does not imply Y_{Yes} as $x_4, x_7 \in Y_{Yes}$ but $x_3 \in Y_{No}$. Thus Table 1 is inconsistent as $f(x_4, d) \neq f(x_3, d)$ and $f(x_7, d) \neq f(x_3, d)$.

To represent an inconsistent decision table, the CRSA establishes an upper and lower approximation for each decision class, Y. The lower approximation is comprised of all objects that definitely belong to Y, while the upper approximation includes all objects that possibly belong to Y. It can be said that an object x_i definitely belongs to a concept Y if $[x_i]_C \subseteq Y$ and that x_i possibly belongs to a concept Y if $[x_i]_C \cap Y \neq \emptyset$. Thus, the lower and upper approximations are defined as follows:

$$\underline{R}_B(Y) = \{x \in U : [x]_B \subseteq Y\} = \bigcup \{[x]_B : [x]_B \subseteq Y\}$$

$$\overline{R}_B(Y) = \{x \in U : [x]_B \cap Y \neq \emptyset\} = \bigcup \{[x]_B : [x]_B \cap Y \neq \emptyset\}$$

$$\overline{R}_B(Y) - \underline{R}_B(Y) = BND_B(Y)$$

The boundary region, $BND_B(Y)$, contains those objects that possibly, but not certainly, belong to Y. Conversely, the set $U - \overline{R}_B(Y)$ is the outside region containing those objects that certainly do not belong to Y. In our example, the lower and upper approximations for Y_{Yes} are $\underline{R}_C(Y_{Yes}) = X_4 = \{x_5, x_6\}$ and $\overline{R}_C(Y_{Yes}) = X_4 \cup X_3 = \{x_3, x_4, x_5, x_6, x_7\}$, and the boundary region contains the objects $BND_B(Y_{Yes}) = \{x_3, x_4, x_7\}$.

Let $F = \{Y_1, Y_2, \ldots, Y_n\}$ represent a classification, i.e. a set of decision classes. The quality of approximation of classification, $\gamma_B(F)$, with respect to attributes B, expresses the ratio of all objects covered by the lower approximation $\underline{R}_B(F) = \{\underline{R}_B(Y_1), \underline{R}_B(Y_2), \ldots, \underline{R}_B(Y_n)\}$ over all objects in U. The quality of approximation is expressed as:

$$\gamma_B(F) = \frac{\sum_{t=1}^{n} \left|\underline{R}_B(Y_t)\right|}{|U|}$$

Dominance-based rough set approach (DRSA)

Under the DRSA [39] the relations between objects are no longer made by the indiscernibility relation as described in the CRSA [29]. In its place, the DRSA introduces a new dominance relation that allows for ordinal attributes with preference-ordered domains wherein a monotonic relationship exists between the attribute and the decision classes. An example of such a relationship occurs when a "better" or "worse" value of an attribute leads to a "better" or "worse" decision class.

Notation

A decision table in the DRSA is expressed in the same way as the CRSA. To differentiate between attributes with and without a preference-ordered domain, those with a preference order are called criteria while those without are referred to as attributes, as in the CRSA.

In the DRSA the domain of criteria $a \in A$ is completely preordered by the outranking relation $\succeq_a$, representing the preference order of the domain. The outranking relation is also applicable for comparing two objects such that for $x_i, x_j \in U$, $x_i \succeq_a x_j$ means that x_i is at least as good as (outranks) x_j with respect to the criterion $a \in A$.

Commonly, the domain of a criteria a is a subset of real numbers, $V_a \subseteq R$ and the outranking relation is then a simple order "$\geq$" on real numbers such that the following relation holds: $x_i \succeq_a x_j \Leftrightarrow f(x_i, a) \geq f(x_j, a)$. This relation is straightforward for gain-type criteria (the more, the better), and can be easily reversed for cost-type criteria (the less, the better).

Using Table 1 as an example, the decision criterion d is preference-ordered such that a positive diagnosis of coronary disease is assumed to be the "preferred" decision class. Criterion preference relations are then organized in the direction of the decision class; values which generally contribute to the incidence of coronary disease are preferred over those which indicate lower risk, much in the same way that a positive diagnosis indicates presence of coronary disease. For the criteria in Table 1, higher values are preferred to lower values—as in the case of *Age*, *SystBP*, and *HDL*—and "Yes" is preferred to "No"—as in the case of *Smoker* and *Diabetic*. No such preference relation exists for *Gender*; as such, it is considered an attribute.

Let $T = \{1, \ldots, n\}$ represent increasing indexes corresponding to the order of preferences of the decision criterion d. Then, the decision table is partitioned into n classes Y_t, $t \in T$, where each object $x \in U$ is assigned to one and only one class Y_t. The decision classes are

preference-ordered according to the decision maker, i.e. for all $r, s \in T$ such that for $r > s$ the objects from class Y_r are strictly preferred to the objects from class Y_s.

For our example in Table 1, $Y_1 = \{x_1, x_2, x_3\}$ corresponds to patients without a coronary disease and $Y_2 = \{x_4, x_5, x_6, x_7\}$ corresponds to the patients with a coronary disease. Therefore, each patient in Y_2 is preferred to each patient in Y_1.

Set approximations

In the DRSA, the approximated sets are upwards and downwards unions of decision classes rather than individual decision classes as in the CRSA. Upward and downward unions of classes are defined as:

$$Y_t^{\geq} = \bigcup_{s \geq t} Y_s \quad \text{and} \quad Y_t^{\leq} = \bigcup_{s \leq t} Y_s, \; s, t \in T$$

For any pair of objects $(x_i, x_j) \in U$, x_i dominates x_j with respect to a set of condition attributes $P \subseteq C$, denoted by $x_i D_P x_j$, if the following conditions are satisfied simultaneously:

$$\begin{aligned} x_i &\succeq_q x_j, \quad \text{for all critera } \; q \in P \\ f(x_i, a) &= f(x_j, a), \; \text{for all attributes } \; a \in P \end{aligned}$$

The dominance relation defines two sets called dominance cones, where for each $x_i \in U$:

$$\begin{aligned} D_P^+(x_i) &= \{x_j \in U : x_j D_P x_i\}, \text{ representing the set of objects that dominates } x_i \\ D_P^-(x_i) &= \{x_j \in U : x_i D_P x_j\}, \text{ representing the set of objects dominated by } x_i \end{aligned}$$

Considering the dominance cones, the lower and upper approximations of the union of decision classes are defined as follows. The lower approximation $\underline{R}_P\left(Y_t^{\geq}\right)$ represents objects that certainly belong to $Y_t^{\geq}$, such that there is no other object that dominates x and belongs to a decision class inferior to Y_t. Similarly, the lower approximation $\underline{R}_P\left(Y_t^{\leq}\right)$ represents objects that certainly belong to $Y_t^{\leq}$, with no other object dominated by x and belonging to a decision class superior to Y_t. The upper approximations represent objects that possibly belong to one of the upward or downward unions of decision classes.

$$\begin{aligned} \underline{R}_P\left(Y_t^{\geq}\right) &= \left\{x \in U : D_P^+(x) \subseteq Y_t^{\geq}\right\} \\ \overline{R}_P\left(Y_t^{\geq}\right) &= \bigcup_{x \in Y_t^{\geq}} D_P^+(x) = \left\{x \in U : D_P^-(x) \cap Y_t^{\leq} \neq \emptyset\right\} \\ \underline{R}_P\left(Y_t^{\leq}\right) &= \left\{x \in U : D_P^-(x) \subseteq Y_t^{\leq}\right\} \\ \overline{R}_P\left(Y_t^{\leq}\right) &= \bigcup_{x \in Y_t^{\leq}} D_P^-(x) = \left\{x \in U : D_P^+(x) \cap Y_t^{\geq} \neq \emptyset\right\} \end{aligned} \tag{1}$$

Similar to the CRSA, the boundary regions are defined as:

$$\begin{aligned} BND_P Y_t^{\geq} &= \overline{R}_P\left(Y_t^{\geq}\right) - \underline{R}_P\left(Y_t^{\geq}\right) \\ BND_P Y_t^{\leq} &= \overline{R}_P\left(Y_t^{\leq}\right) - \underline{R}_P\left(Y_t^{\leq}\right) \end{aligned}$$

Using our example decision table, Table 1, and considering the full set of condition attributes, it can be seen that $x_4 \; D_C \; x_3$, and furthermore $D_C^+(x_4) = \{x_3, x_4, x_7\}$, $D_C^-(x_4) = \{x_3, x_4, x_7\}$. Considering the dominance cones for all patients, the lower and upper approximations of the union of decision classes are $\underline{R}_C\left(Y_2^{\geq}\right) = \{x_5, x_6\}$, $\overline{R}_C\left(Y_2^{\geq}\right) = \{x_3, x_4, x_5, x_6, x_7\}$, $\underline{R}_C\left(Y_1^{\leq}\right) = \{x_1, x_2\}$, $\overline{R}_C\left(Y_1^{\leq}\right) = \{x_1, x_2, x_3, x_4, x_7\}$ and the boundary regions are $BND_C Y_2^{\geq} = BND_C Y_1^{\leq} = \{x_3, x_4, x_7\}$.

For every subset of attributes $P \subseteq C$, the quality of approximation of the decision classes Y with respect to the attributes P, $\gamma_P(Y)$, is defined as the proportion among all objects in U of objects consistently defined with respect to the attributes P and the decision classes Y.

$$\gamma_P(Y) = \frac{\left|U - \left\{\bigcup_{t \in T} BND_P Y_t^{\leq}\right\}\right|}{|U|} = \frac{\left|U - \left\{\bigcup_{t \in T} BND_P Y_t^{\geq}\right\}\right|}{|U|}$$

The variable consistency DRSA

The variable consistency DRSA (VC-DRSA) allows the decision maker to relax the strictness of the dominance relation, thus accepting a limited number of inconsistent objects in the lower approximation, according to an object consistency level threshold, $l \in (0, 1]$. In practice, by selecting this consistency level l, a patient $x \in U$ becomes a member of the lower approximation of a given upward union if at least $l * 100$ % of the patients dominating x also belong to that decision class. By allowing inconsistencies, the VC-DRSA avoids over fitting the training set and thus may be more effective in classifying new cases.

The lower approximations of the VC-DRSA-based model are represented as follows:

$$\begin{aligned} \underline{R}_P^l\left(Y_t^{\geq}\right) &= \left\{x \in Y_t^{\geq} : \frac{|D_P^+(x) \cap Y_t^{\geq}|}{|D_P^+(x)|} \geq l\right\} \\ \underline{R}_P^l\left(Y_t^{\leq}\right) &= \left\{x \in Y_t^{\leq} : \frac{|D_P^-(x) \cap Y_t^{\leq}|}{|D_P^-(x)|} \geq l\right\} \end{aligned}$$

Continuing with the example described in Table 1, setting $l = 0.6$ moves the objects x_4 and x_7, previously included in the upper approximation $\overline{R}_C\left(Y_2^{\geq}\right)$, to the lower approximation of class $Y_2^{\geq}$, i.e: $\underline{R}_C^{0.6}\left(Y_2^{\geq}\right) = \{x_4, x_5, x_6, x_7\}$. This follows from $\frac{|D_C^+(x_i) \cap Y_t^{\geq}|}{|D_C^+(x_i)|} = \frac{2}{3} \geq l$, for $i = 4, 5, 6, 7$.

Decision rules

There are a number of methods available for induction of decision rules from the lower or upper approximations of the decision classes [40–42] or from reducts extracted from the decision table [43]. Decision rules in this study were obtained using the MODLEM [40, 41] and VC-DomLEM [42] algorithms for the induction of classical and dominance-based rough set rules, respectively. In both cases, decision rules are induced from approximations of decision classes. Both the MODLEM and VC-DomLEM algorithms generate a minimal set of decision rules using a minimal number of rule conditions, thus the inclusion of MODLEM allows for an evaluation of the impact of accounting for the preference order information in the VC-DRSA. Once decision rules have been induced, the collection of these rules can then be used to classify unseen objects—in the case of our example table, a new patient who may have cardiac disease.

A decision rule has the form *if A then B*, or $A \rightarrow B$, where A is called the antecedent and B the consequent of the rule. The antecedent is a logical conjunction of descriptors and the consequent is the decision class or union of decision classes suggested by the rule.

Formally, in the CRSA, decision rules are generated from the lower or upper approximations. For example, for an approximation containing objects with descriptors r with respect to a set of condition attributes, $B_r \subseteq C$, a decision rule is expressed as

$$if \bigwedge_i \left(f(x, a_i) = r_{a_i}\right) \; then \; x \in Y_t$$

where $a_i \in B_r$ is an attribute found in the attribute set B_r, and $r_{a_i} \in V_{a_i}$ and Y_t are the attribute values and a decision class, respectively, of the objects in the rule-generating approximation. From our example in Table 1, a decision rule induced with MODLEM from the lower approximation $\underline{R}_{\{Age,Smoker\}}(Y_{Yes}) = \{x_5, x_6\}$ would be: if *Age* = H and *Smoker* = Yes then *Coronary Disease* = Yes.

In the DRSA, decision rules are induced from the lower approximations and the boundaries of the union of decision classes. From the lower approximations, two types of decision rules are considered. Decision rules generated from the P-lower approximation of the upward union of decision classes $Y_t^{\geq}$ are described by

$$if \left(\bigwedge_i \left(f(x, b_i) \geq r_{b_i}\right)\right) \bigwedge \left(\bigwedge_j \left(f(x, a_j) = r_{a_j}\right)\right) \; then \; x \in Y_t^{\geq}$$

where $b_i \in P$ are criteria, $a_j \in P$ are attributes, $r_{b_i} \in V_{b_i}$ and $r_{a_j} \in V_{a_j}$. From the example in Table 1, the P-lower approximation of the upward union of the decision class, $Y_2^{\geq} = \underline{R_C} Y_2 = \{x_5, x_6\}$, leads to the following decision rule:

- *If* Gender = M *and* Age $\geq$ H *and* HDL $\geq$ L *and* Diabetic $\geq$ Yes *and* Smoker = Yes, *then* Coronary Disease = Yes.

Decision rules generated from the P-lower approximation of the downward union of classes $Y_t^{\leq}$ are described by

$$if \left(\bigwedge_i \left(f(x, b_i) \leq r_{b_i}\right)\right) \bigwedge \left(\bigwedge_j \left(f(x, a_j) = r_{a_j}\right)\right) \; then \; x \in Y_t^{\leq}$$

where $b_i \in P$ are criteria, $a_j \in P$ are attributes, $r_{b_i} \in V_{b_i}$ and $r_{a_j} \in V_{a_j}$. From the example in Table 1, the P-lower approximation of the downward union of classes $Y_1^{\leq} = \underline{R_C} Y_1 = \{x_1, x_2\}$, leads to the following decision rules:

- *If* Gender = F *and* Age $\leq$ H *and* SystBP $\leq$ M *and* HDL $\leq$ L *and* Diabetic $\leq$ No *and* Smoker $\leq$ Yes, *then* Coronary Disease = No
- *If* Gender = M *and* Age $\leq$ H *and* SystBP $\leq$ M and HDL $\leq$ L *and* Diabetic $\leq$ No *and* Smoker $\leq$ Yes, *then* Coronary Disease = No

The boundaries $BND_P Y_t^{\geq}$ and $BND_P Y_t^{\leq}$ generate the following rules

$$if \left(\bigwedge_i \left(f(x, b_i) \geq r_{b_i}\right)\right) \bigwedge \left(\bigwedge_j \left(f(x, b_j) \leq r_{b_j}\right)\right)$$
$$\bigwedge \left(\bigwedge_k \left(f(x, a_k) = r_{a_k}\right)\right) \; then \; x \in Y_t \cup Y_{t+1} \cup \ldots \cup Y_s$$

where $b_i, b_j \in P$ are criteria, $a_k \in P$ are attributes, $r_{b_i} \in V_{b_i}$, $r_{b_j} \in V_{b_j}$ and $r_{a_k} \in V_{a_k}$ (note i and j are not necessarily different). From the example in Table 1, the boundary decision classes $BNDY_2^{\geq} = BNDY_1^{\leq} = \{x_3, x_4, x_7\}$, leads to the following decision rule:

- *If* Age $\geq$ M *and* SystBP $\geq$ M *and* HDL $\geq$ H *and* Diabetic $\geq$ No *and* Smoker $\geq$ No *and* Age $\leq$ M *and* SystBP $\leq$ M *and* HDL $\leq$ H *and* Diabetic $\leq$ No *and* Smoker $\leq$ No *and* Gender = F, *then* Coronary Disease = (No, Yes)

The MODLEM and the VC-DomLEM algorithms utilize a heuristic strategy called *sequential covering* [44] to iteratively construct a minimal set of minimal decision rules. The sequential covering strategy successively constructs a set of decision rules for each upward and downward union of decision classes in a training set by selecting, at each iteration, the "best" decision rule, after which the training objects described by the rule conditions are removed. Subsequent iterations again select the best decision rule and remove the covered objects until reaching a stopping criteria or until all of the objects in the

unions of decision classes are described by a rule in the rule set.

To ensure minimality, antecedent descriptors, called elementary conditions, of each rule are checked at each iteration and redundant elementary conditions are removed. Similarly, redundant rules are removed from the final rule set.

In both algorithms, decision rules are grown by consecutively adding the best available elementary condition to the rule. CRSA elementary conditions are evaluated in the MODLEM algorithm in terms of either the class entropy measure [45] or Laplacian accuracy [46]; the former was used in this study. MODLEM does not restrict elementary conditions to those attributes not currently in the rule; as such, multiple elementary conditions may contain the same attribute. Therefore, a decision rule induced by MODLEM may contain antecedents in which attribute values are described as belonging to a range or a set of values or as being greater or less than a particular value.

Dominance-based elementary conditions are evaluated according to a rule consistency measure. VC-DomLEM provides three such measures; the rule consistency measure used in this study is μ, as described in [47]. For the sake of clarity, Y_t shall be used to represent an individual decision class in the CRSA or alternatively an upward or downward union of decision classes, $Y_t^{\geq}$ or $Y_t^{\leq}$, with respect to the DRSA. The consistency, μ, of a proposed rule, r_{Y_t}, suggesting assignment to Y_t is defined as

$$\mu(r_{Y_t}) = \frac{\left|\left[\Phi(r_{Y_t})\right] \cap \underline{Y}_t\right|}{\left|\left[\Phi(r_{Y_t})\right]\right|}.$$

Here $\left[\Phi(r_{Y_t})\right]$ indicates the set of objects described by the elementary conditions in r_{Y_t}. The elementary condition, *ec*, that is selected for inclusion is that which leads to the highest rule consistency measure $\mu(r_{Y_t} \cup ec)$ when combined with the current set of elementary conditions in the proposed rule. In the event of a tie, the elementary condition providing greatest coverage of the new rule is selected, by $\left|\left[\Phi(r_{Y_t} \cup ec)\right] \cap \underline{Y}_t\right|$. The rule consistency measure, μ, was also implemented in MODLEM to relax consistency requirements and to allow more general rules to be induced. For further details on the MODLEM and VC-DomLEM algorithms, the reader is referred to [40–42, 47].

To classify an unseen object, a standard voting process [43] is used to allow all rules to participate in the decision process, arriving at a patient classification by majority vote. Each rule is characterized by two support metrics. The left hand side (LHS) support is the number of patients in the table whose attributes match the antecedent, i.e: $|[\Phi(r)]|$, while the right hand side (RHS) support indicates the number of patients matching both the antecedent and the consequent of the rule, i.e: $|[\Phi(r)] \cap Y_t|$. For a new, unseen patient, any rule whose antecedent descriptors match the patient descriptors "fires" by contributing as votes the RHS support for each decision class. For example, drawing up the example Table 1, the decision rule *If* Age = H *and* Smoker = Yes, *then* Coronary Disease = Yes has LHS = 2 since its antecedent matches patient x_5 and x_6 and RHS = 2 since its antecedent and consequent match the same patients. A new patient matching the antecedent of this rule will receive two votes for decision class *Yes* and zero votes for decision class *No*.

Once all rules have "voted", the number of votes for each decision class is normalized against the total number of LHS support for all fired rules. The resultant ratio of RHS to LHS support is considered a frequency-based estimate of the probability that the patient belongs to the given decision class.

A final classification is therefore determined according to a threshold value, $\tau \in [0, 1]$. A patient is classified as not surviving six months if the estimated probability of death in six months is greater than τ. In the event of an estimated probability equal to τ, or in the absence of any fired rules (no rule matches the patient profile), classification is not possible and the patient is labeled *undefined*. For example, if the threshold value is set as 0.5 and the voting process yields an estimated probability of 70 %, then the patient is classified as not surviving the six month period.

Dataset description

SUPPORT dataset

The dataset used in this study is the SUPPORT (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments) prognostic model dataset [48], a study of 9,105 terminally ill patients. SUPPORT enrolled patients, 18 years or older, who met specific criteria for one of nine serious illnesses, who survived more than 48 hours but were not discharged within 72 hours. Patients were followed such that survival and functional status were known for 180 days after entry. The result of the SUPPORT study is a prognostic model for 180-day survival estimation of seriously ill hospitalized adults based on cubic splines and a Cox regression model. Given the inclusion criteria (described in full in Appendix 1 of [12]), the dataset is ideal for the present research in regards to clinical applicability, completeness of data, and comparability of results.

We consider as condition attributes the variables used in the SUPPORT prognostic model equation [12] to ensure consistency. The SUPPORT variables include ten physiologic variables in addition to the diagnosis groups, age, number of days in the hospital before entering the study, presence of cancer, and neurologic function as recorded in the SUPPORT data. Attribute names, descriptions and value ranges are listed in Table 2.

Table 2 Description of attributes from SUPPORT dataset

Variable name	Description	Patient distribution		
Numerical condition attributes		Range	Mean	Std. Dev.
age	Age of the patient	18–101	62.65	15.59
alb	Serum albumin	0.4–29	2.95	0.87
bili	Bilirubin	0.1–63	2.55	5.32
crea	Serum creatinine	0.09–21.5	1.77	1.69
hday	Number of days in hospital at study entry	1–148	1.00	9.13
hrt	Heart rate	0–300	97.16	31.56
meanbp	Mean arterial blood pressure	0–195	84.55	27.70
pafi	Blood gasses, $PaO_2/(.01 * FiO2)$	12–890.4	239.50	109.70
resp	Respiration rate	0–90	23.33	9.57
scoma	SUPPORT coma score, based on Glasgow coma scale	0–100	12.06	24.63
sod	Sodium	110–181	137.60	6.03
temp	Temperature in °C	31.7–41.7	37.10	1.25
wblc	White blood cell count	0.05–200	12.35	9.27
Categorical condition attributes		Patients	Percentage (%)	
dzgroup	Diagnosis group:			
	ARF/MOSF w. sepsis	3,513	38.59	
	CHF	1,387	15.23	
	Cirrhosis	508	5.56	
	Colon cancer	512	5.62	
	Coma	596	6.54	
	COPD	967	10.60	
	Lung cancer	908	9.97	
	MOSF w. malignancy	712	7.81	
ca	Presence of cancer:			
	Yes	1,252	13.75	
	No	5,993	65.84	
	Metastasis	1,858	20.40	
Decision attribute		Patients	Percentage (%)	
d.6months	Death occurred within 6 months:			
	Yes	4,263	46.83	
	No	4,840	53.17	

Values of 0 for *hrt*, *meanbp* and *resp* correspond to cardiac arrests during the day when the measurements were taken

The median survival time for the patients in the study is 223 days.

Figure 1 shows the patients Kaplan-Meier survival curve with respect to number of days until death. The SUPPORT study inclusion criteria was designed to include patients with 50 % risk of death at 180 days; as seen in Table 2 death prior to 180 days was observed in approximately 47 % of patients.

General observations regarding the influence of condition attributes can be made by analyzing their relation in the proportion of patients surviving the six month period. For example, the Kaplan-Meier survival curve in Fig. 2 shows that a significant portion (75 %) of patients with coma or multi-organ system failure with malignancy (MOSF w/ malig) do not survive longer than 180 days, but patients with congestive heart failure (CHF) or chronic obstructive pulmonary disease (COPD) tend to live longer than 180 days.

Data preprocessing

In its published form, the SUPPORT dataset contains 9,105 cases. Missing physiological attribute values are

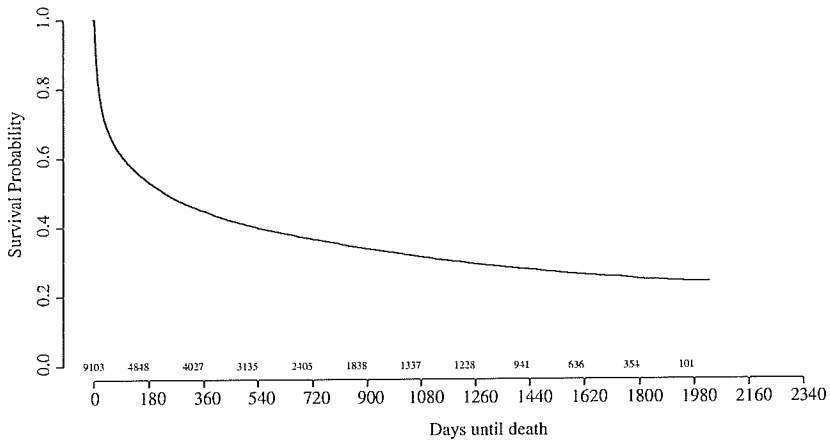


Fig. 1 Kaplan-Meier survivability of patients with respect to number of days until death

filled in with a standard fill-in value representing a normal physiological response, as provided by the SUPPORT authors in [48]. It is also worth noting that in the SUPPORT study, where neurologic functioning of the patient is recorded in the SUPPORT coma score (*scoma*), a patient for whom it was not possible to establish a Glasgow coma score was given a *scoma* value of zero. After missing data imputation, two cases have missing values in physiological attributes not addressed in the SUPPORT data set. The two incomplete cases were removed and the remaining 9,103 cases were considered in the development of the prognostic models.

Discretization

Discretization is the process by which appropriate categorical ranges are found for variables with a continuous value range. There are a number of methods available for unsupervised discretization that operate without

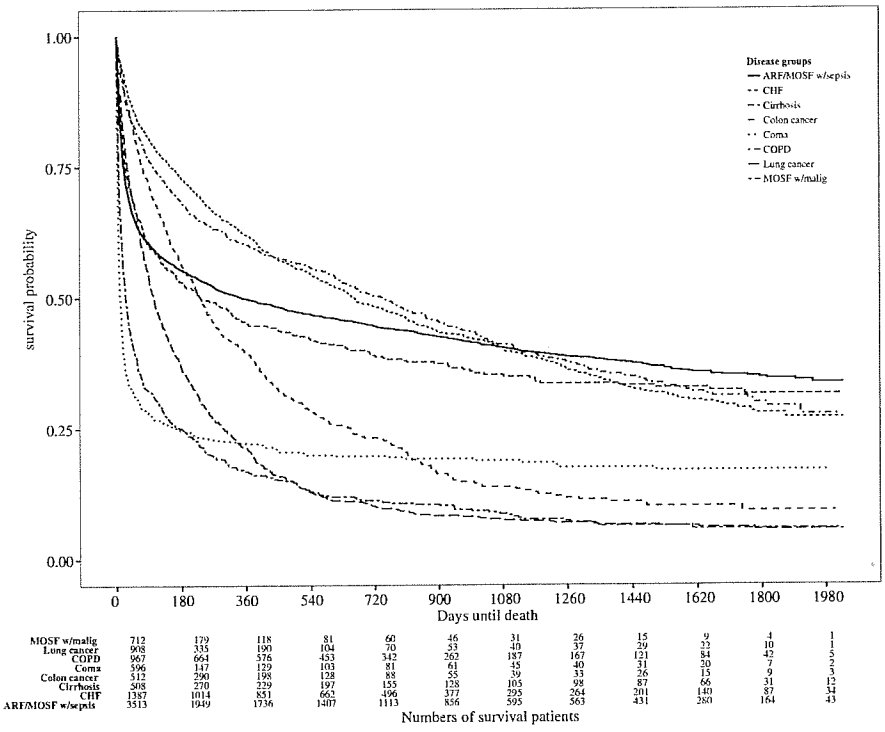


Fig. 2 Kaplan-Meier survival curve by *dzgroup*

input from the decision maker and are based only on the information available in the data table. In this work, however, discretization was primarily performed using the Acute Physiology and Chronic Health Evaluation (APACHE) III scoring system [11], a clinically accepted scoring system designed to estimate the risk of death in ICU patients. In this sense, the use of the APACHE III scoring system represents a research-validated, clinically appropriate, expert discretization scheme. This choice is founded on the proposition that expert discretization via APACHE III will result in medically and contextually relevant classification rules and data collection requirements, thus increasing the accessibility of the proposed prognostic model and ensures directly comparable rule sets for all evaluated rule-based methods.

APACHE III scores are designed to increase monotonically with respect to risk of death and thus provide the necessary preference relations for the DRSA. APACHE III scores for any given variable are close to zero for normal or only slightly abnormal values of that variable and increase according to increased severity of disease. For example, normal pulse rates of 50–99 bpm are given a score of 0, while elevated and lowered levels, 100–109 and 40–49 bpm respectively, are both given a score of 5. Thus, higher APACHE III scores are preferred to lower scores, as the higher scores indicate greater severity of disease and therefore greater risk of death within six months (considered the positive diagnosis). Discretization is not a requirement of any of the methods used in this study, however the APACHE III scores provide the monotonic preference relations for the DRSA and are used for the all of the rule-based methods.

For the rule-based methods considered in this study, the nine physiologic variables and the age variable were transformed to their representative APACHE III scores. The remaining physiologic variables not included in APACHE III—neurologic function, *scoma*, and blood gasses, *pafi*—were discretized using clinically accepted categorizations [49, 50]. The variable *hday* was discretized using the boolean reasoning algorithm [43]. Table 3 shows the categories defined in this process. Higher values of each of these variables are preferred to lower values.

Table 3 Discretized attributes not in APACHE III

Attribute	Description	Categorization
scoma	Minor	$(*, 9]$
	Moderate	$(9, 44]$
	Severe	$(44, *)$
pafi	Normal	$[300, *)$
	Severe defect in gas exchange	$[200, 300)$
	Acute respiratory distress syndrome	$[0, 200)$
hday	Short	$(*, 44]$
	Long	$(44, *]$

Experimental design

This section provides details on the implementation and performance evaluation procedures for the comparison of the classification methods used in this study. The following two sections, describe the RSA and comparative methods respectively, the software used for their implementation and the selection of appropriate parameters for each of the methods. Finally, the methods for performance evaluation are discussed.

The general schema of the experimental design is as follows: after selecting appropriate parameters for each of the methods, 5-fold cross validation was used to divide the data into training and testing sets. Methods with decision rule outputs were trained and tested on the discretized data set to demonstrate expected performance of a clinically credible rule set. Methods without decision rule outputs were trained on the raw, non-discretized, data set. For these methods, designed to be applied to continuous variables, discretization does not improve clinical credibility and would likely hinder performance [51, 52].

Rough set rule induction and classification

MODLEM algorithm for CRSA decision rules CRSA decision rules were obtained using the MODLEM algorithm as described in [40] and [41], implemented by the authors in the R programming language [53]. Decision rules were generated from the lower approximations with a rule consistency level $\mu \geq m$. The rule syntax follows the presentation in section Decision rules.

VC-DomLEM algorithm for VC-DRSA decision rules Dominance-based rules were obtained using the VC-DRSA as described in section The variable consistency DRSA and the VC-DomLEM algorithm as implemented in jMAF [54]. VC-DomLEM decision rules were generated from the lower approximation of each decision class, with an object consistency level threshold l. The syntax of the VC-DRSA decision rules is as shown in section Decision rules. Only decision rules with rule consistency measure μ greater than the rule consistency threshold l are included in the classification model. Note that the rule consistency threshold and the object consistency threshold are equal and set at l.

Parameter selection In order to select the most appropriate models for comparison, the performance of the rough set based models was evaluated for varying levels of rule consistency, m and l, for the CRSA and VC-DRSA respectively. Classifier performance at a particular value of m or l is dataset-dependent; however, in general, values

close to one provide rule sets that are more conservative in describing the training set objects, while values closer to zero provide rule sets that are more general. Thus, to find the appropriate balance between strict, descriptive models that are prone to overfitting and overly general models that provide little useful information, the RSA models were evaluated at $m, l = 0.1, 0.2, 0.4, 0.6, 0.8, 1.0$.

Comparative methods

To evaluate the performance of the RSA-based prognostic models, logistic regression, SVM, and RF were applied to the non-discretized SUPPORT dataset. To ensure directly comparable rule sets, C4.5 was applied to the discretized SUPPORT dataset. Each of these methodologies was applied using the software package Weka 3.6.9 [55], within which appropriate parameters were selected for SVM, C4.5 and RF using GridSearch with 10-fold cross validation settings. Logistic regression was selected for its popularity in classification models using non-censored data and in clinical settings [18, 56].

Support vector machines, originally presented in [57], find separating boundaries between decision classes after input vectors are non-linearly mapped into a high dimensional feature space. Support vector machines have been investigated in survival analysis applications [58] as they—similar to the RSA-based methods—automatically incorporate non-linearities and do not make *a priori* assumptions about factor interactions. SVM-based models are known to perform well at classification tasks, however they do not provide clinician-interpretable justification for their results [59]. Support vector machines were selected to evaluate whether the increased accessibility of the RSA-based methods involves a trade-off in accuracy.

C4.5 is a well known algorithm for generating a decision tree using information entropy to select the best splitting criteria at each node [60]. A decision tree built by C4.5 can be expressed as a set of if-then decision rules, thus providing a comparative decision rule based method. Decision trees were obtained using the Weka J48 implementation [60] of the C4.5 algorithm.

Random forests is a popular ensemble classification method based on decision trees [61]. The random forests algorithm builds an ensemble of decision trees, where each tree is built on bootstrap samples of training data with a randomly selected subset of factors.

Performance evaluation methods

The performance of the models was tested by measuring the discriminatory power of both the m- and l-consistent decision rules sets when applied to the reserved testing data. For our notation, a classification of $d.6months = Yes$ is referred to as a positive classification, and $d.6months = No$ is negative. Sensitivity is defined as the fraction of patients who did not survive six months and are correctly classified by the model, or the fraction of true positive classifications of all test patients who did not survive six months. Conversely, specificity is defined as the fraction of patients who did survive six months and were correctly classified by the model, or the fraction of true negatives of all test patients who did survive six months.

The overall accuracy of the classification models is reported in terms of area under the receiver operating characteristic (ROC) curve, or AUC (area under the curve). The ROC curve graphs the sensitivity of the classifier, or the true positive rate, versus 1 − specificity, the false positive rate, as the threshold probability, τ, for positive classification is varied from 0 to 1. The best overall classification performance is realized when AUC is equal to 1, while an AUC of 0.5 indicates a classifier performance no better than random selection. Best separation between decision classes is realized at the threshold corresponding to the point on the ROC curve closest to the point (0, 1).

In order to select the most appropriate MODLEM and VC-DomLEM-based models for comparison, two performance issues related to the generated rule set were considered: coverage and AUC of the model. The coverage of the classification model is defined as the percentage of testing set patients for whom a classification is possible. Additionally, to evaluate the number of rules that would fire for an unseen patient, we collected information on the number of rules matching each test case patient for the evaluated levels of m and l.

Cohen's Kappa coefficient was computed for both the selected RSA-based models and the comparative models [62]. Cohen's Kappa coefficient is designed to measure the agreement between two classification methods, but it is commonly used to measure model performance by comparing a classifier with a random allocation of patients among the decision classes. A value of zero indicates classification accuracy equivalent to chance (zero disagreement).

Performance of the prognostic models was evaluated using a 5-fold cross validation procedure [63] wherein training and testing sets are repeatedly selected. Cross validation is a well known method that provides a reasonable estimate of the generalization error of a prediction model. In 5-fold cross validation, the entire dataset is randomly divided into five subsets, or folds, and then each fold (20 % of the dataset) is used once as a testing set, with the remaining folds (80 %) used for training.

Results

This section presents the results obtained using MODLEM, VC-DomLEM, logistic regression, SVM, C4.5 and RF models for six-month life expectancy prognostication of terminally ill patients. The results are analyzed and compared.

In order to select appropriate m and l values for MODLEM and VC-DomLEM-based models, respectively, the performance of these models was evaluated first. AUC and coverage for each evaluated m and l level are shown in Table 4. Figures 3 and 4 display the number of rules that fire for each patient in the five testing folds for each m and l value. Based on these results, $m = l = 0.6$ was chosen as the rule consistency parameter for both algorithms for further evaluation with the comparative methods.

The quality of approximation is 0.9244 for the CRSA, 0.3110 for the DRSA and finally 0.9014 for the VC-DRSA where the object consistency parameter $l = 0.6$.

Table 5 describes the number of rules and the number of descriptors in each rule for the two rough set approach-based classifiers at the selected consistency level of 0.6. The average number of MODLEM decision rules in the five rule sets generated by cross validation is 773 rules, with mean and maximum length of 3.65 and 8 descriptors, respectively. In Fig. 3, it can be seen that at rule consistency levels of $m = 0.2$ and $m = 0.1$, the number of rules fired per patient is always 2. This is because the rule set is generated by only two attributes and each rule contains only one attribute in the antecedent. The VC-DomLEM decision rules are on average slightly longer, with mean and maximum length of 6.85 and 13 elementary conditions, respectively. The mean total number of VC-DomLEM rules is 1,095 rules.

For SVM, the gamma (γ) and cost parameter (C) were evaluated between 10^{-1} and 10^{5} at increments of 10^{-1}; final selected parameters were $\gamma = 0.1$ and $C = 100$. For RF, the number of trees was explored between 10 and 1,000 trees at intervals of 10; the optimal number of trees thus obtained was 500. The maximum number of attributes selected at each bootstrap iteration was also explored in the range of 1 to 15 attributes, with best performance observed when the number of attributes was limited to 1. In the case of C4.5, the confidence factor used for pruning was evaluated between 0.1 and 0.9 with increments of 0.1 and 0.5 was selected. The minimum number of instances per leaf for the C4.5 decision tree was explored in steps of 1 between 1 and 100, with best performance achieved with a minimum of 40 instances per leaf. The pruned C4.5 trees contained an average of 74 nodes over the 5 cross validation folds.

Table 4 AUC and coverage for MODLEM and VC-DomLEM algorithms with l and m-consistent rules

	MODLEM		VC-DomLEM	
m, l	AUC	Coverage (%)	AUC	Coverage (%)
0.1	0.6646	100.00	0.7280	99.88
0.2	0.6646	100.00	0.7279	99.87
0.4	0.6888	100.00	0.7277	99.65
0.6	0.6974	97.41	0.7173	98.72
0.8	0.6419	86.72	0.7093	76.85
1.0	0.6158	80.08	0.6559	35.89

The performance of all of the evaluated classification models is shown in Table 6, where Cohen's kappa coefficient [62] and AUC are reported for each classifier, averaged over the 5 cross validation folds. Highest average kappa coefficient was achieved by RF with $\bar{\kappa} = 0.37$. Second highest average kappa coefficient was achieved by VC-DomLEM, logistic regression and SVM at $\bar{\kappa} = 0.35$. The MODLEMand C4.5 classifiers achieved $\bar{\kappa} = 0.32$ and 0.31, respectively. Average sensitivity and specificity for each of the models are also shown in Table 6. For each model and cross validation fold configuration, the sensitivity and specificity were recorded at the threshold at which both values are simultaneously maximized. This threshold is equivalent to the point on the ROC plot closest to the upper left corner and represents the point of maximum accuracy of the model.

Discussion

All of the methodologies show fair classification accuracy given that Kappa coefficients are in the range of 0.20 to 0.40 [64]. The results presented in Table 6 show that all of the methods have similar AUC with the best performing algorithm being RF (AUC = 0.7459) and the worst being MODLEM (AUC = 0.6974). The best performing method among the decision- and rule-based methods was VC-DomLEM with an average AUC of 0.7173.

With respect to MODLEM and VC-DomLEM, m and l are clearly critical values in determining model performance. Together, Table 4 and Figs. 3 and 4 demonstrate that selecting $m = l = 0.6$ balances the accuracy and coverage achieved by the rough set based classifiers against the amount of inconsistency allowed in each. In the case of MODLEM, $m = 0.6$ is associated with highest AUC and acceptable coverage. However, in the case of VC-DomLEM, reducing l below 0.6 provides only marginal benefits in terms of AUC and coverage but greatly increases the amount of inconsistency allowed in the generated rules.

The quality of approximation for the CRSA classifier is 0.9244. The difference between the quality of approximation in the CRSA and the DRSA is the inclusion of the preference-ordering information, determined by the APACHE III scores. In the case of the DRSA, a strict application of this information in determining the lower approximation leads to few patients in the lower approximations, thus reducing the overall quality of approximation. Consequently, decision rules generated from this approximation are too specific and less suitable for generalizing to the classification of new cases. It is therefore reasonable to relax the conditions for assignment

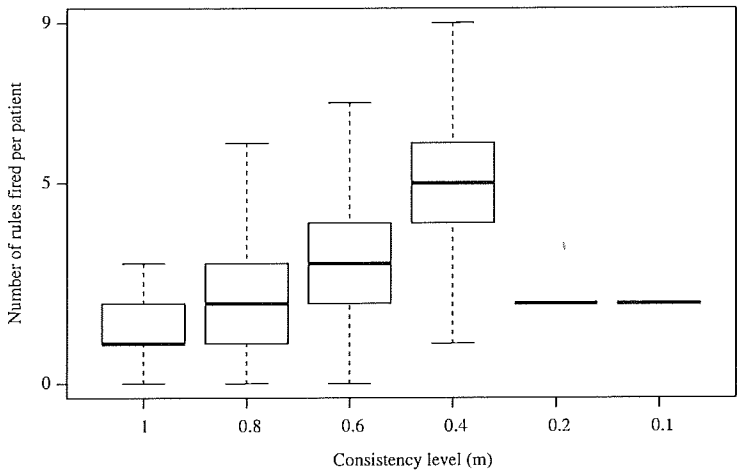


Fig. 3 Number of rules fired in each test case for *m*-consistent MODLEM classifiers

of objects to lower approximations. Thus, using the VC-DRSA and setting the object consistency parameter $l = 0.6$, results in an improved quality of approximation of 0.9014.

All of the rule- or decision-tree-based methods demonstrated somewhat reduced performance when compared with the non-rule based classifiers. The worst-performing rule-based method, MODLEM, had an AUC 0.049 below the best-performing method, RF (0.6974, MODLEM, vs. 0.7459, RF). In contrast, VC-DomLEM demonstrated average AUC much closer to that of RF, with an average difference of only 0.029 (0.7173, VC-DomLem, vs. 0.7459, RF). In practice, this relatively small difference in performance is likely to be outweighed by the accessibility of the rule-based format of VC-DomLEM, while such benefits would be less justified in the case of MODLEM.

Interpretation and usability of decision rules

Clinical credibility in prognostic models depends in part on the ease with which physicians and patients can understand and interpret the results of the models, in addition to the accuracy of the information they provide. The RSA-based prognostic models present the physician with a list of matched decision rules, offering significant advantages by increasing both the traceability of the model and the amount of information included in its results. This advantage is further increased in the case of VC-DomLEM, where dominance-based decision rules permit greater information density per rule by including attribute value ranges in each rule.

Table 7 contains the decision rules that fire for an example patient selected from the SUPPORT data set. This patient was 41 years old with a primary diagnosis of coma.

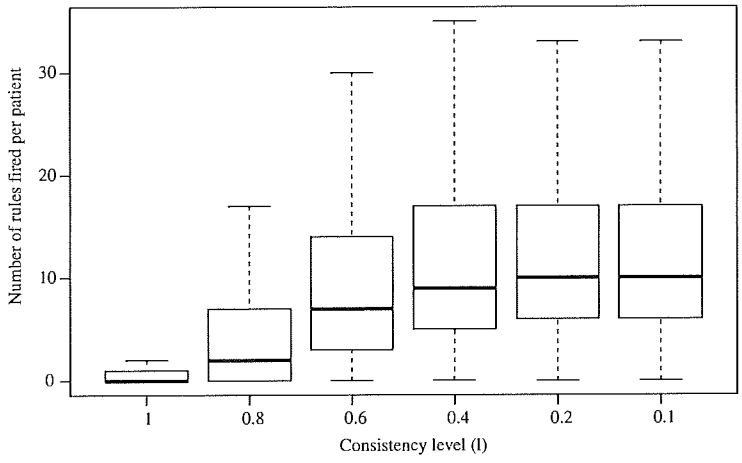


Fig. 4 Number of rules fired in each test case for *l*-consistent VC-DomLEM classifiers

Table 5 Number of descriptors and rules in MODLEM and VC-DomLEM induced decision rule sets, for $m = I = 0.6$ consistent rules, across the five cross validation folds

Method	Mean number of rules	Descriptors in rules		
		Min.	Max.	Mean
MODLEM	773	1	8	3.65
VC-DomLEM	1095	2	13	6.85

The patient displayed moderate head injury on the Glasgow Coma Scale, elevated levels of creatinine (1.60 mg/dL) and respiratory rate (26 bpm), normal levels of sodium (133 mEq/L), low white blood cell count (1.90 cells/nL) and mean blood pressure of 107 bpm. Both the MODLEM and VC-DomLEM classifiers correctly predict that the patient will not survive six months (the patient in fact survived only 4 days).

The VC-DomLEM classifier predicts *d.6months* = *Yes* with an associated score of 80 %, based on the two rules (Rules 5 and 6). As can be seen in Table 7, Rule 5 isolates the combination of Coma and elevated creatinine and sodium levels as a key predictor of six-month survival. In the case of Rule 5, 51 patients in the training set have similar conditions as the example patient, of which 47 did not survive six months. On the other hand, Rule 6 somewhat counterbalances this prediction, pointing to 8 young patients with moderate coma who have been in the hospital less than 44 days, of whom all 8 survived six months.

The MODLEM classifier provides a less specific prediction, classifying the example patient as not surviving six months with an associated score of 55 %. Upon further investigation, the rules matching the example patient (Rules 1–4) are more general than the rules provided by the VC-DomLEM classifier. Rules 1–3 provide general rules that point to the age, level of head trauma and primary diagnosis of the patient. Considering only these three rules, the associated score would be *d.6months* = *Yes* with a score of 54 %, but this score is revised slightly by Rule 4 further in favor of *d.6months* = *Yes*. Rule 4 isolates normal average heart beat, high respiratory rate and low (and also very high) white blood cell counts.

For all of the classifiers, a final prediction and associated score are presented by the classifier. However, only in the case of MODLEM and VC-DomLEM is the prediction further supported by the set of rules from which said prediction derived. Thus, the gestalt survival expectation is presented without loss of contradictory information, providing the physician with both the prognostication as well as supporting and contradicting information. In contrast, while a decision tree obtained using C4.5 can be represented as a set of rules, only a single rule representing the matching terminal node is returned to the physician. Among the rule-based methods, those rules derived from the VC-DRSA naturally include attribute value ranges for which the rule is valid, succinctly providing even more information to the physician and further increasing the utility of the life expectancy prediction. In a clinical setting, this set of rules serves to support clinical decisions for future treatment or palliative care strategies as well as to support the explanation of these decisions to the involved patient and their family.

Decision tree models offer the additional benefit of visually representing the entire model in a single structure, and given their hierarchical structure can be used to guide the decision process of the physician [65]. Decision trees models are most useful when built with the input of domain experts as pruning visually complex decision trees must balance tradeoffs between accuracy and simplicity [66]. Many methods exist for the visualization of decision trees and the performance of visually-tuned decision trees may be comparable to more complex versions of the same model [67].

A further benefit of the rule-based methods is that the rules clearly indicate the patient characteristics most relevant to their survival expectation. This increases the transparency and interpretability of the classification process, strengthening the accessibility, and hence credibility, of the model. Additionally, the decision rules do not individually involve all of the condition attributes. This offers the advantage of providing potentially acceptable results should a particular prognostic factor be difficult or too costly to ascertain for a patient [34].

Table 6 Summary of performance evaluation results of the classification models, averaged over the 5 cross validation folds, with standard deviations

Method	AUC	Kappa	Sensitivity	Specificity	Threshold (τ)
VC-DomLEM	0.7173 (0.014)	0.35 (0.03)	0.6391 (0.042)	0.7175 (0.033)	0.4234 (0.045)
MODLEM	0.6974 (0.015)	0.32 (0.03)	0.6447 (0.038)	0.6862 (0.037)	0.4597 (0.042)
C4.5	0.7088 (0.018)	0.31 (0.04)	0.6078 (0.055)	0.7254 (0.070)	0.4531 (0.095)
RF	0.7459 (0.014)	0.37 (0.02)	0.6384 (0.044)	0.7388 (0.039)	0.4872 (0.022)
Log. Reg.	0.7421 (0.009)	0.35 (0.01)	0.6374 (0.055)	0.7282 (0.058)	0.4715 (0.050)
SVM	0.7352 (0.009)	0.35 (0.02)	0.6526 (0.050)	0.7132 (0.040)	0.4056 (0.034)

Table 7 Selected decision rules from the CRSA using MODLEM and the VC-DRSA using VC-DomLEM

			RHS	
	CRSA rules using MODLEM	LHS	*d.6months = No*	*d.6months = Yes*
1.	If *age_score*[a] = 0	969	593 (61 %)	376 (39 %)
2.	If *scoma* = Moderate	1016	399 (39 %)	617 (61 %)
3.	If *dzgroup* = Coma	465	119 (26 %)	346 (74 %)
4.	If *hrt_score*[b] = 0 AND *resp_score*[c] = 6 AND *wbc_score*[d] = 5	47	11 (23 %)	36 (77 %)
	VC-DRSA rules using VC-DomLEM			
5.	If *dzgroup* = Coma AND *crea_score*[e] ≥ 4 AND *sod_score*[f] ≥ 2	51	4 (8 %)	47 (92 %)
6.	If *dzgroup* = Coma AND *scoma* ≤ Moderate AND *hday* ≤ Short AND *age_ score*[a] ≤ 0	8	8 (100 %)	0 (0 %)

[a] *age_score:* 0 = (*age* ≤ 44)
[b] *hrt_score:* 0 = (50 ≤ *hrt* ≤ 99)
[c] *resp_score:* 6 = (25 ≤ *resp* ≤ 34)
[d] *wbc_score:* 5 = ((1 ≤ *wbc* ≤ 2.9) or (*wbc* ≥ 25))
[e] *crea_score:* ≥ 4 = (*crea* ≥ 1.5)
[f] *sod_score:* ≥ 2 = ((*sod* ≤ 134) or (*sod* ≥ 155))

This is in stark contrast to SVM, neural networks, and other black-box methods where very little insight is available to a decision maker as to how an outcome was predicted. While similar performance in terms of accuracy was seen for all of the classification models, the rule-based models naturally express results in terms of a set of decision rules, a benefit that is not present in logistic regression, RF, or the mentioned black-box methods. As an ensemble method, the RF method functionally reduces to a black-box style model, despite its use of decision trees.

Decision analysis for hospice referral

Consider the costs—economic, emotional and physical—associated with the decision to enter hospice care. These costs are justified for patients who either enter hospice care at the appropriate time or for those who do not enter hospice care when they could benefit from curative treatment. These cases represent true positive and true negative classifications. A higher emotional and physical cost is born by patients sent to hospice care but who ultimately survive six months—a false positive. The highest cost of all, emotionally, economically and physically is born by the patient and his or her family when costly treatment is prolonged for a patient who should have been referred to a hospice care program—a false negative. In this last case, some or all of the benefits of hospice care would be lost while the stresses and economic burden of aggressive treatment are endured.

In this light, the threshold parameter, τ (described in section Decision rules), can be seen as a representation of the patient and family's preference for hospice care treatment and their risk tolerance for a mistaken referral. The threshold parameter relates sensitivity to specificity and stipulates the required level of certainty for a positive classification. A higher threshold value requires a higher probability of not surviving six months for the classification of a patient as a hospice candidate, decreasing the sensitivity and increasing specificity (indicating a preference for continued treatment). Conversely, a lower threshold value increases sensitivity while reducing specificity, indicating a preference for avoiding the costly mistake of unnecessary treatment.

As this threshold value is a subjective matter and varies between physicians, patients and family members, one suggested approach [68] involves the measurement of the amount of regret the decision maker would have should an incorrect decision be made. As medical decisions must take into account the preferences of those ultimately affected by the decision, this application of regret theory allows for the formal treatment of those preferences by calculating the threshold value as a function of the measured anticipated regret.

Conclusions

This paper contributes to the growing body of research in RST—and its extensions—as a prognostic modeling framework and highlights the strengths of this approach in terms of accessibility. The non-rule-based methods—RF, logistic regression, and SVM—were found to more accurately predict death within six months, however the benefits of the rule-based methods may outweigh the

performance differential, particularly in the case of VC-DomLEM where this difference was small. The intuitive structure of the rough set approaches, built on indiscernibility and dominance relations and expressed in terms of if-then decision rules, offers both more insight into the modeling process and more opportunity for the knowledge extraction process to incorporate the personal preferences of those making and being affected by the decision.

The performance of the classifiers presented in this study is good but sub-optimal, indicative of a challenging problem in need of further research. The increased performance achieved by the variable consistency approach suggests a dataset of highly diverse patients. Future research will explore methods to improve the overall classifier performance and address this diversity by building localized models for patient subgroups using rough sets concepts to group patients with similar differentiating characteristics.

A recent study developed a six-month survival prognostic model primarily based on the Medicare Health Outcomes Survey responses of community-dwelling elderly patients [69]. This model, named the Patient-Reported Outcome Mortality Prediction Tool (PROMPT), achieved comparable AUC using only basic medical information, indicating that the performance of classification models for six-month survival is still a major issue for the targeted domain of hospice referral recommendation.

An important limitation of this study is that patient-specific disease progression over time is not considered, in part due to the static nature of the data set used. Future research must address the temporal aspect of disease progression, a consideration often missing in other prognostic models for hospice referral. The progression of a terminal illness is often highly non-linear by nature and generally does not present as a steady decline over time but rather as periods of relative stability marked by turning points of acute decline. A prognostic model that takes into account this temporal aspect may possibly provide both more accurate life expectancy prognoses and more useful information for palliative care planning.

Availability of supporting data

The data set supporting the results of this article is publicly available at http://biostat.mc.vanderbilt.edu/wiki/Main/SupportDesc.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

EG and GA are the principal authors; EG performed the data analysis and implemented the different algorithms. EG and GA wrote the manuscript. GA and AY participated in writing/editing and critically revising the manuscript with substantial contributions. AY, AT, BD and LB contributed to the interpretation of the data and revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Department of Army under grant # W81 XWH-09-2-0175.

The authors thank Dr. Jurek Blaszczynski from the Laboratory of Intelligent Decision Support Systems at the Poznan University of Technology, Poznan, Poland, for his collaboration in providing valuable information about the Dominance Based Rough Set Approach.

Author details

[1]Threshold Tuning Optimization Team Citigroup, 3800 Citibank Center, Tampa, 33610 FL, USA. [2]Department of Industrial and Management Systems Engineering, University of South Florida, 4202 E. Fowler Avenue, ENB 118, Tampa, FL 33620, USA. [3]Department of Systems and Information Engineering, University of Virginia, 151 Engineer's Way, Charlottesville, VA 22904, USA. [4]Department of Hematology and Health Outcomes and Behavior, H. Lee Moffitt Cancer Center & Research Institute, University of South Florida, 12902 Magnolia Drive, Tampa, FL 33612, USA.

Received: 8 January 2015 Accepted: 3 November 2015
Published online: 25 November 2015

References

1. Kane RL, Klein SJ, Bernstein L, Rothenberg R, Wales J. Hospice role in alleviating the emotional stress of terminal patients and their families. Medical Care. 1985;23(3):189–97.
2. Bulkin W, Lukashok H. Rx for dying: The case for hospice. N Engl J Med. 1988;318(6):376–8.
3. Dawson NJ. Need satisfaction in terminal care settings. Soc Sci Med. 1991;32(1):83–7.
4. Christakis NA. Timing of referral of terminally ill patients to an outpatient hospice. J Gen Intern Med. 1994;9(6):314–20.
5. Teno JM, Shu JE, Casarett D, Spence C, Rhodes R, Connor S. Timing of referral to hospice and quality of care: length of stay and bereaved family members' perceptions of the timing of hospice referral. J Pain Symptom Manag. 2007;34(2):120–5.
6. Christakis NA, Escarce JJ. Survival of medicare patients after enrollment in hospice programs. N Engl J Med. 1996;335(3):172–8.
7. Lee KL, Pryor DB, Harrell FE, Califf RM, Behar VS, Floyd WL, et al. Predicting outcome in coronary disease statistical models versus expert clinicians. Am J Med. 1986;80(4):553–60.
8. Wyatt JC, Altman DG. Commentary: Prognostic models: clinically useful or quickly forgotten? BMJ. 1995;311(7019):1539–41.
9. Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. Science. 1989;243(4899):1668–74.
10. Beck JR, Pauker SG, Gottlieb JE, Klein K, Kassirer JP. A convenient approximation of life expectancy (the "deale") ii: use in medical decision-making. Am J Med. 1982;73(6):889–97.
11. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 1991;100(6):1619–36.
12. Knaus WA, Harrell FE, Lynn J, Goldman L, Phillips RS, Connors AF, et al. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. Ann Intern Med. 1995;122(3):191–203.
13. Christakis NA. Predicting patient survival before and after hospice enrollment. Hosp J. 1998;13:71–88.
14. Gripp S, Moeller S, Bölke E, Schmitt G, Matuschek C, Asgari So. Survival prediction in terminally ill cancer patients by clinical estimates, laboratory tests, and self-rated anxiety and depression. J Clin Oncol. 2007;25(22):3313–20.
15. Glare P, Sinclair C, Downing M, Stone P, Maltoni M, Vigano A. Predicting survival in patients with advanced disease. Eur J Cancer. 2008;44(8):1146–56.
16. Hyodo I, Morita T, Adachi I, Shima Y, Yoshizawa A, Hiraga K. Development of a predicting tool for survival of terminally ill cancer patients. Jpn J Clin Oncol. 2010;40(5):442–8.

17. Han PKJ, Lee M, Reeve BB, Mariotto AB, Wang Z, Hays RD, et al. Development of a prognostic model for six-month mortality in older adults with declining health. J Pain Symptom Manag. 2012;43(3):527–39.
18. Hosmer DW, Lemeshow S, Inc N. Applied Survival Analysis: Regression Modeling of Time to Event Data. New Jersey, USA: John Wiley & Sons, Inc; 2008, p. 404.
19. Dybowski R, Gant V, Weller P, Chang R. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. The Lancet. 1996;347(9009):1146–50.
20. Jayasurya K, Fung G, Yu S, Dehing-Oberije C, De Ruysscher D, Hope A, et al. Comparison of bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. Med Phys. 2010;37(4):1401–7.
21. Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. Crit Care Med. 2001;29(2):291–296.
22. Segal MR. Features of tree-structured survival analysis. Epidemiology. 1997;8(4):344–6.
23. Hwang SS, Scott CB, Chang VT, Cogswell J, Srinivas S, Kasimis B. Prediction of survival for advanced cancer patients by recursive partitioning analysis: role of karnofsky performance status, quality of life and symptom distress. Cancer Invest. 2004;22(5):678–87.
24. Bazan J, Osmólski A, Skowron A, Ślęzak D, Szczuka M, Wroblewski J. Rough set approach to the survival analysis In: Alpigini J, Peters J, Skowron A, Zhong N, editors. Rough Sets and Current Trends in Computing. Lecture Notes in Computer Science, vol. 2475. Malvern, PA, USA: Springer; 2002. p. 522–529.
25. Pattaraintakorn P, Cercone N, Naruedomkul K. Hybrid rough sets intelligent system architecture for survival analysis In: Peters JF, Skowron A, Marek VW, Orłowska E, Słowinski R, Ziarko W, editors. Transactions on Rough Sets VII. Berlin Heidelberg: Springer; 2007. p. 206–24.
26. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005;34(2): 113–27.
27. Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med. 2002;26(1-2):1–24.
28. Simons P. Critical notice of timothy williamson, nagueness. Int J Philos Stud. 1996;4:321–7.
29. Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data. Norwell, MA: Springer; 1992.
30. Pawlak Z. Vagueness a rough set view In: Mycielski J, Rozenberg G, Salomaa A, editors. Structures in Logic and Computer Science, vol. 1261. Springer Berlin Heidelberg; 1997. p. 106–117. doi:10.1007/3-540-63246-8_7.
31. Pattaraintakorn P, Cercone N. Integrating rough set theory and medical applications. Appl Math Lett. 2008;21(4):400–3.
32. Hart A, Wyatt J. Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. Inform Health Soc Care. 1990;15(3):229–36.
33. Tsumoto S. Modelling medical diagnostic rules based on rough sets In: Polkowski L, Skowron A, editors. Rough Sets and Current Trends in Computing. Lecture Notes in Computer Science, vol. 1424. Berlin Heidelberg: Springer; 1998. p. 475–82. doi:10.1007/3-540-69115-4_65.
34. Komorowski J, Øhrn A. Modelling prognostic power of cardiac tests using rough sets. Artif Intell Med. 1999;15(2):167–91.
35. Paszek P, Wakulicz-Deja A. Applying rough set theory to medical diagnosing In: Kryszkiewicz M, Peters J, Rybinski H, Skowron A, editors. Rough Sets and Intelligent Systems Paradigms. Lecture Notes in Computer Science, vol. 4585. Berlin Heidelberg: Springer; 2007. p. 427–35.
36. Ningler M, Stockmanns G, Schneider G, Kochs HD, Kochs E. Adapted variable precision rough set approach for EEG analysis. Artif Intell Med. 2009;47(3):239–61.
37. Long-Jun H, Li-pin D, Cai-Ying Z. Prognosis system for lung cancer based on rough set theory. In: Proceedings of the 2010 Third International Conference on Information and Computing, vol. 4. Washington, DC, USA: IEEE Computer Society; 2010. p. 7–10.
38. Son CS, Kim YN, Kim HS, Park HS, Kim MS. Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches. J Biomed Inform. 2012;5(45):999–1008.
39. Greco S, Matarazzo B, Slowinski R. Rough sets theory for multicriteria decision analysis. Eur J Oper Res. 2001;129(1):1–47.
40. Stefanowski J. The rough set based rule induction technique for classification problems. In: Proceedings of 6th European Congress on Intelligent Techniques and Soft Computing EUFIT 98. Aachen: IEEE Computer Society; 1998. p. 109–113.
41. Stefanowski J. On combined classifiers, rule induction and rough sets In: Peters J, Skowron A, Düntsch I, Grzymała-Busse J, Orłowska E, Polkowski L, editors. Transactions on Rough Sets VI, Lecture Notes in Computer Science, vol. 4374. Springer; 2007. p. 329–50.
42. Blaszczyński J, Słowiński R, Szelag M. Probabilistic rough set approaches to ordinal classification with monotonicity constraints In: Hüllermeier E, Kruse R, Hoffmann F, editors. Computational Intelligence for Knowledge-Based Systems Design. Lecture Notes in Computer Science, vol. 6178. Berlin Heidelberg: Springer; 2010. p. 99–108.
43. Bazan JG, Nguyen HS, Nguyen SH, Synak P, Wróblewski J. Rough set algorithms in classification problem In: Polkowski L, Tsumoto S, Lin T, editors. Rough Set Methods and Applications. Studies in Fuzziness and Soft Computing, vol. 56. Heidelberg, Germany: Physica-Verlag GmbH; 2000. p. 49–88.
44. Han J. Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2005.
45. Grzymala-Busse JW, Stefanowski J. Three discretization methods for rule induction. Int J Intell Syst. 2001;16(1):29–38.
46. Clark P, Boswell R. Rule induction with cn2: Some recent improvements In: Kodratoff Y, editor. Machine Learning – EWSL-91. Lecture Notes in Computer Science, vol. 482. Berlin Heidelberg: Springer; 1991. p. 151–63.
47. Blaszczynski J, Slowinski R, Szelag M. Sequential covering rule induction algorithm for variable consistency rough set approaches. Inf Sci. 2011;181(5):987–1002.
48. Harrell FE. SUPPORT Datasets. 2010. http://biostat.mc.vanderbilt.edu/wiki/Main/SupportDesc. Accessed 17 Nov 2015.
49. Stein SC. Minor head injury: 13 is an unlucky number. The Journal of Trauma and Acute Care Surgery. 2001;50(4):759–60.
50. Martin L. Reviews, notes, and listings: pulmonary medicine: All you really need to know to interpret arterial blood gases. Ann Intern Med. 1993;8(118):656.
51. Cohen J. The cost of dichotomization. Appl Psychol Meas. 1983;7(3): 249–53.
52. Van Belle G. Statistical Rules of Thumb vol. 699. Hoboken, NJ, USA: John Wiley & Sons; 2011.
53. Team RC. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.R-project.org/. Accessed 17 Nov 2015.
54. Błaszczyński J, Greco S, Matarazzo B, Słowiński R, Szelag M. jmaf - dominance-based rough set data analysis framework In: Skowron A, Suraj Z, editors. Rough Sets and Intelligent Systems - Professor Zdzisław Pawlak in Memoriam. Intelligent Systems Reference Library, vol. 42. Springer; 2013. p. 185–209.
55. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol. 2011;2(3):1–27.
56. Bender R, Grouven U. Ordinal logistic regression in medical research. J R Coll Physicians Lond. 1997;31(5):546–51.
57. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3): 273–97.
58. Van Belle V, Pelckmans K, Van Huffel S, Suykens JA. Support vector methods for survival analysis: a comparison between ranking and regression approaches. Artificial Intelligence In Medicine. 2011;53(2): 107–18.
59. Barakat N, Bradley AP. Rule extraction from support vector machines: a review. Neurocomputing. 2010;74(1-3):178–90.
60. Quinlan JR. C4. 5: Programs for Machine Learning vol. 1. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1993.
61. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
62. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
63. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1995. p. 1137–1143.

64. Ben-David A. Comparison of classification accuracy using Cohen's weighted kappa. Expert Systems with Applications. 2008;34(2):825–32.
65. Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: an overview and their use in medicine. J Med Syst. 2002;26(5):445–463.
66. Bohanec M, Bratko I. Trading accuracy for simplicity in decision trees. Mach Learn. 1994;15(3):223–50.
67. Stiglic G, Kocbek S, Pernek I, Kokol P. Comprehensive decision tree models in bioinformatics. PLoS ONE. 2012;7(3):33812.
68. Tsalatsanis A, Hoz I, Andrew V, Djulbegovic B, Hozo I, Vickers A, Djulbegovic B. A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making. BMC Med Inform Decis Mak. 2010;10(51):51.
69. Han PKJ, Lee M, Reeve BB, Mariotto AB, Wang Z, Hays RD, et al. Development of a prognostic model for six-month mortality in older adults with declining health. J Pain Symptom Manag. 2011;43(3):527–39.

Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

Determining optimal threshold for statins prescribing: individualization of statins treatment for primary prevention of cardiovascular disease

Benjamin Djulbegovic MD PhD,[1,3,4] Athanasios Tsalatsanis PhD[2] and Iztok Hozo PhD[5]

[1]Distinguished Professor and Associate Dean, [2]Associate Professor, USF Health Program for Comparative Effectiveness Research, Division for Evidence-Based Medicine, Department of Internal Medicine, University of South Florida, Tampa, FL, USA
[3]Distinguished Professor and Associate Dean, Departments of Hematology and Health Outcome Behavior, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA
[4]Distinguished Professor and Associate Dean, Tampa General Hospital, Tampa, FL, USA
[5]Professor and Chair, Department of Mathematics, Indiana University, Gary, IN, USA

Keywords
clinical guidelines, diagnosis, evaluation

Correspondence
Professor Benjamin Djulbegovic
12901 Bruce B. Downs Blvd, MDC02
Tampa, FL 33612
USA
E-mail: bdjulbeg@health.usf.edu

Accepted for publication: 22 September 2015

doi:10.1111/jep.12473

Abstract

Rationale, aims and objectives The American College of Cardiology and American Heart Association (ACC/AHA) statin guidelines recommend that people with risk of cardiovascular disease (CVD) ≥7.5% over 10 years should be treated with statins. This recommendation ignores individual patient CVD risks and preferences. We compared the ACC/AHA guidelines to the following management strategies a) individualized statins treatment based on Framingham Risk Score (FRS), b) treat none, c) treat all.

Methods We employed regret-based decision curve analysis to evaluate the optimal treatment strategy. We used data on 5013 participants from the second generation of the Framingham Heart Study. We assessed regret of each treatment strategy [treat according to FRS vs. treat none vs. treat all] as a function of emotionally felt loss of treatment benefits and incurred treatment harms. We calculated the difference between regret associated with one strategy compared with the other and expressed it as Net Expected Regret Difference (NERD). Two strategies are identical if NERD = 0.

Results Treatment according to ACC/AHA guidelines represents the optimal strategy only if the patient values avoiding heart disease 12 times more than harms related to statins. For values of benefit/harms (B/H) <12, treatment according to FRS represents the optimal strategy. For B/H <3, 'treat none' represents equally acceptable strategy. Adopting a threshold of 10% recommended by other professional organizations would decrease over-treatment by more than 60% without significantly affecting under-treatment.

Conclusion Under most realistic scenarios, individualizing statins treatment, or not recommending statins at all, represents the optimal strategy for primary prevention of heart disease.

Introduction

Cardiovascular disease (CVD) is the leading cause of mortality and morbidity in the United States [1]. Statins are highly effective drugs for both primary and secondary prevention [2,3]. The American College of Cardiology and American Heart Association (ACC/AHA) have developed new recommendations for the use of statins for primary prevention [4]. The guidelines de-emphasize use of laboratory tests, recommending instead that statins be prescribed according to a patient's estimated risk of CVD [4]. Specifically, if the 10-year risk of CVD that includes coronary heart disease and stroke as well as atherosclerotic peripheral arterial disease is ≥7.5%, then ACC/AHA recommends the patient should take statins, otherwise not [4]. Different organizations such as the National Lipid Association (NLA) and UK National Institute for Health and Care Excellence (NICE) also emphasize using a patient's risk as the basis for treatment, but recommend different thresholds (10–15%) [5,6].

None of these organizations explained how they determined the thresholds they chose. Presumably, the thresholds reflected the organizations' view about how patients ought to weigh statins' benefits and harms. If true, this means each organization

has a different view of patient's preferences for benefits and harms. However, the use of any fixed threshold assumes that all patients have the same preferences for benefits and harms of statins, which is not true. People's preferences vary widely, and no single threshold can be appropriate for everyone. Any attempt to use a fixed threshold will cause overuse in some patients (who have a low threshold) and underuse in others (who have a high threshold) [7].

To use statins appropriately, a guideline should take into account not only each patient's risk of CVD, but also his or her personal preferences about the harms and benefits of statins. It is, however, not known if the preference-based combined with the risk-adjusted approach for primary prevention of CVD is, in fact, superior to the guidelines' recommendations to treat everyone with the risk of CVD above certain threshold. In this paper, we set out to compare the effects of treating everyone according to the recent ACC/AHA, NLA and NICE guidelines versus individualizing guidelines by taking each patient's CVD risk profile and preferences into account.

Methods

Data

The Framingham's risk model for prediction of CVD is widely recommended for determination of CVD risk [8]. We obtained the data from the National Heart, Lung and Blood Institute on 5013 participants from the second generation of Framingham cohort. We calculated individual 10-year CVD risk for 4908 patients according to the Framingham's risk score prediction model (FRS) [8]. We excluded 105 patient records due to missing values. Data were extracted from their original source and compiled into a single data file using SAS Enterprise Guide version 6.1 (SAS Institute Inc., Cary, NC, USA). FRS scores were computed using SAS enterprise guide version 6.1. All other analyses were performed using Stata Statistical Software Release 13, Stata Corp., College Station, TX, USA.

Determination of the threshold of CVD risk above which statins should be prescribed

To calculate the threshold we need (1) objective data on statins' treatment effects and (2) to assess how the patients value benefits and harms of statins.

Empirical evidence

We obtained data on statins benefits and harms from the literature, which evaluated the effects of statins for primary prevention of heart disease. Based on meta-analysis of four randomized trials enrolling 35 254 patients, Taylor *et al.* estimated that statins are associated with 35% [95% confidence intervals (CI): 27–42%] relative risk reduction (RRR) in combined fatal and non-fatal CVD, or 1.33% in terms of the absolute risk reduction [=CVD (placebo) – CVD × (1 – RRR) (statins) = 3.8% – 3.8% × (1 – 0.35) = 1.33%] [3]. They found no statistically significant increase in harms with use of statins over placebo except for diabetes. Statins were associated with the absolute increase in incidence of diabetes of 0.4% (95% CI 0.001–0.8%) over placebo [3]. Statins have also been reportedly associated with a number of other harms (such as myalgia, liver test abnormalities, rhabdomyolysis) [9], which could not be precisely quantified in controlled trials [3]. These harms are typically conveyed qualitatively; similarly, many less tangible benefits of statins, such as avoiding hospitalization, are discussed with the patients using non-quantitative language.

Assessment of patients' preferences

People's preferences are a function of their goals; analytic reasoning can help us achieve our goals [10], but, as philosopher David Hume wrote, 'reason is and ought to be the slave of our passions' [11]. Emotions are felt holistically. Therefore, preferences should also be assessed holistically [12]. For that reason, we employed the regret-based version [13,14] of the threshold model [15,16] to calculate the threshold based on the holistic assessment of statins' benefits and harms. This is because theoretically no administration of treatment or ordering of diagnostic tests can be 100% accurate [17,18]. After making a decision, often under conditions of uncertainty, one may anticipate that another alternative would have been preferable. This knowledge may bring a *regret of omission* (due to failing to receive potentially beneficial services), or *regret of commission* (due to receiving unnecessary and potentially harmful health services) [19–21]. Thus, many if not most medical decisions inevitably involve trade-offs of weighing the consequences of failure to benefit (false negatives, regret of omission) versus unnecessary harms (false positives, regret of commission) that cannot be simultaneously improved for any given health intervention [22]. This trade-off can be captured by asking a decision maker (preferably, patient) how many more times he or she would regret of not receiving health intervention (such as statins) which can prevent CVD compared with unnecessary and potentially harmful administration of treatments (such as statins) [14,23]. From this, we can then calculate the risk threshold (p_t) [16] at which a patient is indifferent between taking treatment, or not [14,23,24]

$$p_t = \frac{1}{1 + \frac{Failure\,to\,benefit\,(RgOmission)(B)}{Unnecessary\,harm\,(RgCommission)(H)}} \quad (1)$$

where B and H is regret based, emotionally felt benefits and harms related to regret of omission and commission, respectively. Therefore, Equation (1) directly links CVD risk threshold (p_t) with the patients' preferences about statins' benefits and harms (B/H). According to Equation (1), the patient should accept statins if risk of CVD $>p_t$; conversely if CVD risk $<p_t$, the patient should refuse statins. As stated earlier, the assessment of B/H as captured via regret-based thresholds stipulates holistic evaluation of failure to benefit versus incurring harms across all treatment effects as experienced by individual decision maker.

Ideally, we would elicit the threshold from each decision maker (i.e. a guideline's panel member, physicians, patients) to help with a given treatment decision. However, this may not be necessary because we can model decisions about treatment over the range of the thresholds using decision curve analysis (DCA). DCA represents an extension of the threshold model [15,16] over a range of thresholds (i.e. *over all possible preferences*). It has been formu-

lated based both on expected utility [25–27] and regret theory [14,23]. The main goal of DCA is to compare different management strategies over a range of decision maker's preferences, which in turn are captured by determining the risk thresholds (p_t). In particular, we want to compare the effects of four strategies: (1) 'treat everyone' according to practice guideline (i.e. for predicted risk ≥7.5, 10 and 15% at 10 years) versus (2) treat according to an 'individualized risk-based' approach using Framingham's prediction model but taking into account the patient's preferences for overtreating as opposed to undertreating with statins, versus (3) 'treat none' (observe), versus (4) 'treat all' (regardless of CVD risk).

In this paper, we extend regret-based DCA [14,23] to incorporate both effects (benefits and harms) of treatments and prediction model based on the FRS score. We identify the optimal management strategy as the one associated with the lowest regret as calculated by 'net expected regret difference' [NERD (strategy 1 vs. strategy 2)] [14], that is, differences between expected regret related to one management strategy versus other (see Appendix for details). If NERD = 0, this means that the management strategies are equal. If NERD <0, then first strategy is better as it generates a lower regret. If NERD >0, then second strategy is associated with a lower regret, and thus better.

We are interested in the following comparisons:

$$NERD[Model_x, Model] = [(FP_x - FP)\cdot(1-p) + (TP_x - TP)\cdot p\cdot RRR]\cdot\left(1+\frac{H_{Rx}}{H}\right) + p\cdot FN_x\cdot\frac{1-x}{x} - p\cdot FN\cdot\frac{1-P_t}{P_t} - (FN_x - FN)\cdot p\cdot\frac{H_{Rx}}{H} \quad (2)$$

where $Model_x$ refers to the ACC/AHA/NLA/NICE guidelines to treat at $P_t = x = 0.075 = 7.5\%$ (*or*, 10% and 15% respectively); *Model* – management according to the Framingham's risk score prediction model (FRS); TP – true positives (patients correctly predicted to have CVD by FRS); FP – false positives (patients incorrectly predicted to have CVD by FRS); TN – true negatives (patients correctly classified not to have CVD by FRS); FN – false negatives (patients incorrectly predicted not to have CVD when they actually had it by FRS); Hte – risk of predictive (FRS) testing itself (assumed to be zero in this analysis); H is regret based, emotionally felt harms due to unnecessary treatment; H_{Rx} – objectively assessed outcomes of statin harms [3].

Because $\frac{1-P_t}{P_t} = \frac{B}{H}$, to facilitate interpretation of the results, we present the analysis for the range of the thresholds 0 to 1 (or B/H from 0 to 20). (Note that the effect of Hrx/H cancels out, so we do not have to run a separate analysis for these variables.)

The procedure to calculate the NERD is explained in the Appendix. Sensitivity analysis was performed by varying the CVD threshold and the RRR associated with statins.

Calculation of the extent of overuse and underuse

Given that no management strategy can be perfect [17], from the policy perspective, we are also interested in finding the extent of overuse or underuse for a range of thresholds. Because policy makers often do not know individual patient's preferences, they may be interested in calculating the magnitude of overuse (FPs) or underuse (FN) for the entire range of the thresholds (0–100%) (i.e. over all possible preferences of their constituency). To calculate the average amount of overuse (underuse) of one strategy versus another, we determined the area under the curve for the difference between two strategies and divided it by the corresponding length of threshold interval.

This research has been approved by the USF IRB (Pro00011951).

Results

Table 1 depicts various patient profiles and the associated 10-year CVD risk to provide context to the results presented in this section. Each patient's CVD risk score is calculated using the FRS equation. Note that the methodology presented applies to any patient evaluated for statin treatment and is not limited to the patient profiles presented in Table 1.

Figure 1a shows the comparison between individualized treatment according to the FRS versus treatment according to the ACC/AHA guidelines. Using our regret framework, we estimated the ACC/AHA panel deemed that failure to prevent CVD (regret of omission) is 12.3 times worse than unnecessary life-long treatment and developing diabetes (regret of commission). As it can be seen, regret of individualizing treatment according to the FRS is much

Table 1 Illustration of various patient characteristics and the associated 10-year cardiovascular disease risk (CVD) scores as computed by the Framingham Risk Score (FRS) equation*

ID	Sex	Age	Sys BP	Total cholesterol	HDL cholesterol	Hypert. meds	Smoker	Diabetes	Risk
1	F	45	125	210	40	No	No	No	4.34%
2	M	45	125	210	40	No	No	No	7.37%
3	M	55	120	180	55	No	Yes	No	14.55%
4	F	60	115	180	60	No	No	Yes	8.21%

*According to the ACC/AHA guideline patients 3 and 4 should be treated with statins because their CVD risk exceeds 7.5%. However, treatment according to the ACC/AHA guidelines represents the optimal option only if the patient values avoiding CVD (failure to benefit) 12 times more than harms related to statins (i.e. development of diabetes). For the values of benefit/harms (B/H) <12, individualized treatment according to FRS represents the optimal management strategy. BP, blood pressure; HDL, high-density lipoprotein.

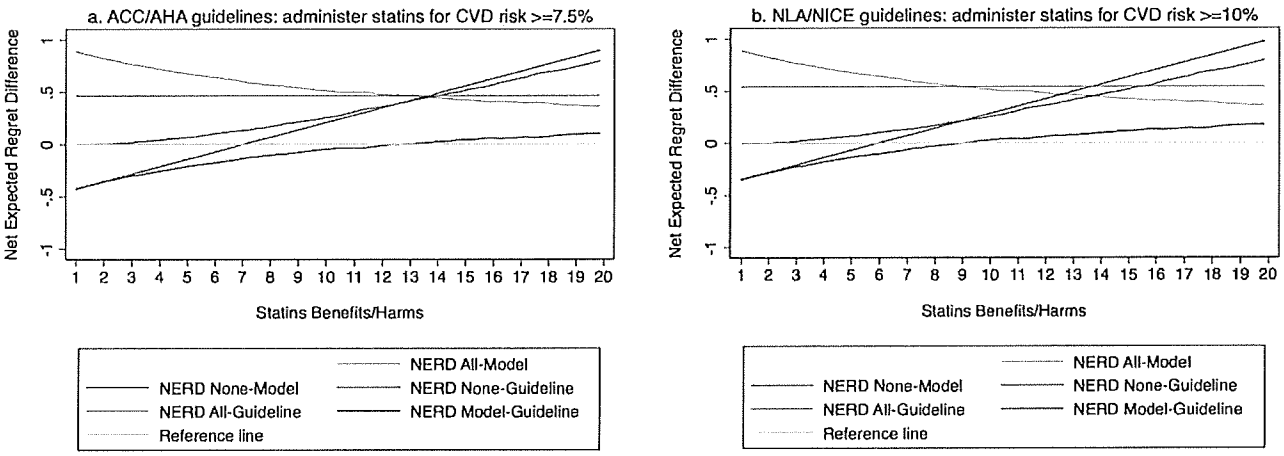


Figure 1 Regret-based decision curve analysis related to decision whether to use statins for primary prevention of cardiovascular disease (CVD). The reference line at 0 indicates no differences between two strategies that are compared. If net expected regret difference (NERD) <0, then the first strategy is better (i.e. associated with the least amount of regret). If NERD >0, then the second strategy is better. The *x*-axis represents the patient's preferences expressed in terms of regret of failing to benefit (B) versus regret of harms (H) associated with statin. 'Model' represent administration of treatment according to Framingham Risk Score (FRS); Guidelines = statin guidelines by American College of Cardiology/American Heart Association (ACC/AHA) (a) and National Lipid Association (NLA) and UK National Institute for Health and Care Excellence (NICE) (b). 'None' means treat no patient with statins. For B/H between 1 and 12, the use of FRS model is the most optimal strategy. The adherence to the ACC/AHA guidelines becomes a superior strategy for the values of B/H >12. Note, however, 'Treat None' is equally acceptable strategy for B/H from 1 to 3.

smaller than regret of acting according to ACC/AHA recommendation [NERD (FRS) < NERD ('treat according to ACC/AHA guidelines')] until B/H = 12.3 (which corresponds to the threshold =7.5% established by the ACC/AHA). Only when the patient deems that failure to prevent CVD (regret of omission) is more than 12 times worse than unnecessary life-long treatment and developing diabetes (regret of commission), ACC/AHA recommendations become better strategy. For B/H <3, 'treat none' was equivalent to the management according to the individualized FRS risk score [NERD ('treat none') – NERD ('treat according to the FRS score' = 0]; thus, for B/H <3, 'treat none' represents equally acceptable management strategy. (Note that 'treat none' was superior strategy to the treatment according to the ACC/AHA guidelines for B/H <6.)

We also compared strategy 'treat all' with all other strategies. This strategy was inferior to all other strategies, and therefore, will not be further discussed.

We also conducted sensitivity analysis by setting the threshold of CVD at 10% according to the NLA and NICE guidelines (Fig. 1b) [5,6]. The NLA and NICE guidelines panels appeared to believe that it is about nine times worse to fail to benefit than to overtreat. That is, the results indicate that treatment according the NLA/NICE guidelines represents the most optimal strategy for B/H >9. For B/H <9, the individualizing treatment according to the FRS score is better strategy and for B/H <3 'treat none' or use of the FRS score is equally acceptable management strategy. Similarly to the ACC/AHA guidelines, 'treat none' was superior strategy to the treatment according to the NLA/NICE guidelines for B/H <6.

In general, higher thresholds set up by the guideline panels will result in the lower B/H ratio at which the patient should accept the guideline recommendations. That is, the higher thresholds will result in lower regret if we fail to administer potentially beneficial treatments; the opposite holds for the lower thresholds (data not shown for the threshold of 15% and higher, and the thresholds lower than 7.5%). Interestingly, the results did not materially change when RRR was varied over all plausible values (0.05–0.5).

Finally, as shown in Fig. 2, the number of unnecessary treatments (FPs) according to ACC/AHA guidelines in comparison with treatment according to FRS over all thresholds (i.e. without reference to individual patients' values) was about 13% and the number of FNs over all thresholds was equal to 3.4%. In other words, adoption of management strategy according to the ACC/AHA guidelines over all plausible patients' values would result in about 13% of more unnecessary treated patients in comparison with individualized treatment according to the FRS score. However, individualized management strategy for all patients would result in under-treatment in about 3% of patients. When we compared the treatment according to the NLA/NICE guidelines versus the individualizing treatment according to the FRS score, the number of FPs (over-treatment if we were to make the recommendation according to the NLA/NICE guidelines) was 8% and the number of FNs (under-treatment if we were adopt to the individualized treatment approach) was 2.7%. That is, adoption of the 10% threshold (instead of 7.5%) would decrease over-treatment by more than 60% without significantly affecting under-treatment.

The comparison of treatment according to the ACC/AHA versus 'treat none' resulted in about 18% of over-treatment according to the guidelines (FPs) at the cost of under-treating about 5% of patients if no patient would take statins for primary prevention of heart disease. The corresponding figures for the NLA/NICE guidelines versus 'treat none' were about 12% (FPs) and about 4% (FNs), respectively.

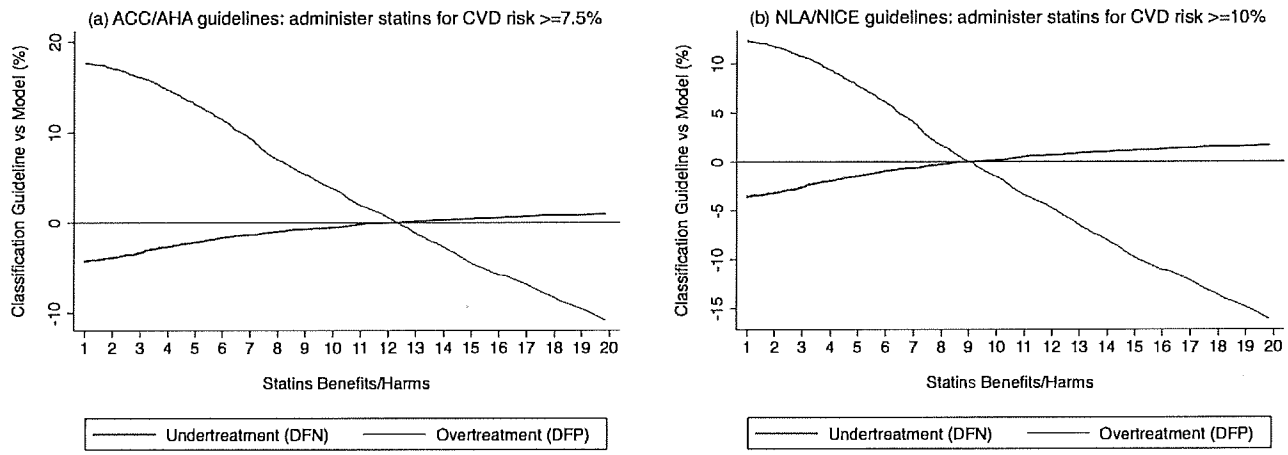


Figure 2 Statin overuse and underuse: comparison of the management according to the guidelines versus individualized risk management according to the Framingham Risk Score (FRS). The reference line at 0 indicates no differences between two strategies that are compared. *y*-Axis indicates a difference in % of accurate classification between two strategies (guidelines – individualized risk management). Red line shows a difference (%) in false positives (DFP) and green line shows a difference (%) in false negatives (DFN). If the value is positive, then the first strategy (i.e. treat according the guidelines) lead to more overuse or underuse, respectively. The opposite is correct if the values are negative – in this case individualized risk approach is associated with more FPs or FNs, respectively. 'Model' represent administration of treatment according to FRS; Guidelines = statin guidelines by American College of Cardiology/American Heart Association (ACC/AHA) (a) and National Lipid Association (NLA) and UK National Institute for Health and Care Excellence (NICE) (b).

Discussion

In this paper, we compared the use of statins for primary prevention of CVD according to the ACC/AHA (and NLA/NICE) guidelines versus individualized risk management according to FRS versus population-oriented approach 'treat all' or 'treat none' of eligible adult patients. Our findings have important policy implications and provide the way how to optimize statin treatment decisions by taking into account individualized CVD risk and patients' preferences.

We found that if the patient values consequences of false negatives (regret of omission)/false positives (regret of commission) >12 times, then her or his values would be aligned with the values of the ACC/AHA guidelines' panel, and consequently, she or he should accept treatment according to the panel's recommendations to receive statins at risk of CVD ≥7.5% . If the patient's values are closer to those of the NLA/NICE guidelines panel, then the patient would accept statins if he or she values false negatives (regret of omission)/false positives (regret of commission) more than nine times.

If the patients' preferences are indeed similar to those of ACC/AHA guidelines panel and statins guidelines are adhered to, it is estimated that 56 million Americans would become eligible to receive statin therapy based on these guidelines [28]. As compared with the prior guidelines, this figure represents an additional 12 million patients in the United States and more than 1 billion people worldwide [29] being considered for statin therapy, with the increase being largely in older adults without CVD [28].

However, the patients' preferences are almost invariably different from the panel's as also evident by the fact that as many as 17% of patients discontinue statin therapy despite their doctors' recommendations [30,31]. Thus, the use of statins is a quintessential preference-based decision [32]. The ACC/AHA (and the NLA/NICE) guidelines are silent on how they arrived at the ratio of benefit/harms at which all eligible patients should receive statins and provide no explanations why patients should accept the panel's particular weighting of the consequences of false-negative versus false-positive treatments. The ACC/AHA statin guidelines are also silent on the method how should doctors consult the patients' preferences and values regarding statin choices, which is considered not only ethically mandatory but may lead to better patient outcomes [33]. Although there is no agreed-upon method for elicitation of patients' preference, we advocate easy-to-use regret-based dual-visual analogue scale (DVAS) as a method for holistic elicitation of patients' preferences that can be easily implemented in every day practice, and upon which this analysis was based upon [14,23,34,35]. Indeed, a substantial body of literature has demonstrated that people are regret averse, which they are motivated to regulate [19,36]. In addition, anticipation of regret leads to more rational choices and vigilant decision making, satisfying most of the criteria for high-quality decisions [19,36,37].

The method we employed is based on holistic assessment of patients' preferences. It has been previously utilized in modelling decisions in various clinical applications [14,23,34,35] and has been empirically validated in a population of physicians [38]. In fact, both strong theoretical rationale and our ongoing experience indicate that the application of regret-based approaches represents one of the most promising approaches to elicit patients' preferences. The approach is not only appealing for its simplicity, but is based on a well-validated approach in literature related to elicitation of regret. A single scale (ranging from zero regret to 100% regret) was also used by Sorum *et al.* [39], as well as in our previous research [14,23,40]. The novelty of our approach was to link these well-validated scales to the threshold model, which, in

turn, can be used to prescribe the action most consistent with a decision maker's values and preferences. Our DVAS methodology is easy to use, easy to understand, time efficient, ideal for clinical setting dominated by time constraints, while tapping in both emotional and deliberative aspects of decision making. Both our experience and the experience by Sorum and colleagues [39] showed that this assessment of asking people to quantify the regret they feel when making choices mirrors what people themselves consider regret, and is easily comprehended. In our ongoing study assessing the role of regret in the end-of-life setting, 96% of 134 patients gave the impression of fully understanding the regret questions (assessed on Likert scale 1–7) [40]. In fact, research suggests that regardless of how regret is elicited, it appears that any instrument will reliably tap into the regret construct [41,42].

The method (Equation 2) includes both individual decision maker subjective perception of benefits and harms (captured via p_t, or B/H assessment) and objective data on statin efficacy (RRR) and disutilities/regret due to statins harms (Hrx). In this respect, our DCA derivation reflects modern cognitive science views, which increasingly accepts dual-processing explanation of human cognition according to which medical decisions can be truly consistent with patients' values and preferences only if they take into account both intuitive (type 1) and analytical (type 2) cognitive processes [43–45]. Regret as cognitive emotion can serve as a link between type 1 and type 2 processes [45,46]. Thus, a decision maker's response to failing to administer appropriate health intervention (underuse), or unnecessary prescribing it (overuse) can be expressed via regret, a cognitive emotion that encompasses both affective and analytical aspect of decision error [19,36].

Although we propose using the DVAS tool to elicit patients' value in order to compare individualized risk management using FRS versus ACC/AHA recommendations (vs. 'treat none' vs. 'treat all'), in our case that was not actually necessary because we evaluated each of management strategy over all thresholds, that is, over all possible patient's preferences.

Interestingly, in comparison with 'treat none', management according to the guidelines over all thresholds resulted in about 18% of patients being over-treated and 5% under-treated. Thus, 'treat none' may still be most acceptable policy for a substantial number of patients, particularly those who regret failure of benefit over unnecessary treatment at ≤3. This may also represent the most rational strategy for policy makers, particularly in light of empirical data that B/H of statins in primary prevention is about 2.5 [3]. This means that if cognitive processes assessing benefits and harms rely on analytical processes only, as one would expect it at the policy levels, 'treat none' would be by far most rational strategy.

In contrast to policy analysis, decision whether to base our decision on the guidelines' recommendations or the prediction model in clinic must depend on the specific, individual patient's preferences as it is the patient who ultimately suffers from the consequences of misclassifications. It is quite possible that our analysis from the policy and individual perspective agrees to the large extent. Nevertheless, we should note that we actually do not know patients' preferences towards statins. However, we would be surprised if the people's estimates of regret of omission versus commission exceeds 10 as this ratio in the higher stake, end-of-life setting using DVAS method was about 8 [40]. Research on elicitation of patients' preferences related to using statins for primary prevention of CVD should be of high priority.

We should note that our estimate of CVD events was based on the FRS [8], which was based on the Framingham cohort followed from 1998 to 2008. The FRS is widely used in clinical practice and has been advocated by the NLA statin guidelines panel [6]. The ACC/AHA guidelines are based on 'pooled cohort risk' calculator [4]. Although we do not know if the ACC/AHA panel would recommend the same threshold based on the FRS model, the recent external validation of four CVS risk scores in multi-ethnic cohort recruited between 2000 and 2002 and followed for 10 years showed that the FRS overestimate CVD events by 25% in relative terms [47,48]. Even though two methods may not agree on the exact computation of CVD risk, it is less relevant issue if one guideline recommends prescription of statins at 7.5%, whereas others at 10%, or 15% [6]. The key point is that administration of statins should be based on assessment of the patients' preferences. Therefore, the lack of agreement between the existing CVD risk prediction tools is unlikely to significantly affect our main findings: for the most of expected patients' preferences, statins should not be used for primary prevention of CVD.

In conclusion, we showed that the optimal strategy for primary prevention of heart disease regardless of the tool used to detect CVD risk rests on individualization of statin treatment based on patient's preferences. If elicitation of patients' preferences is not possible, then the best strategy is to refrain from recommending statins at all.

Acknowledgement

Research support: DOD grant (#W81 XWH 09-2-0175).

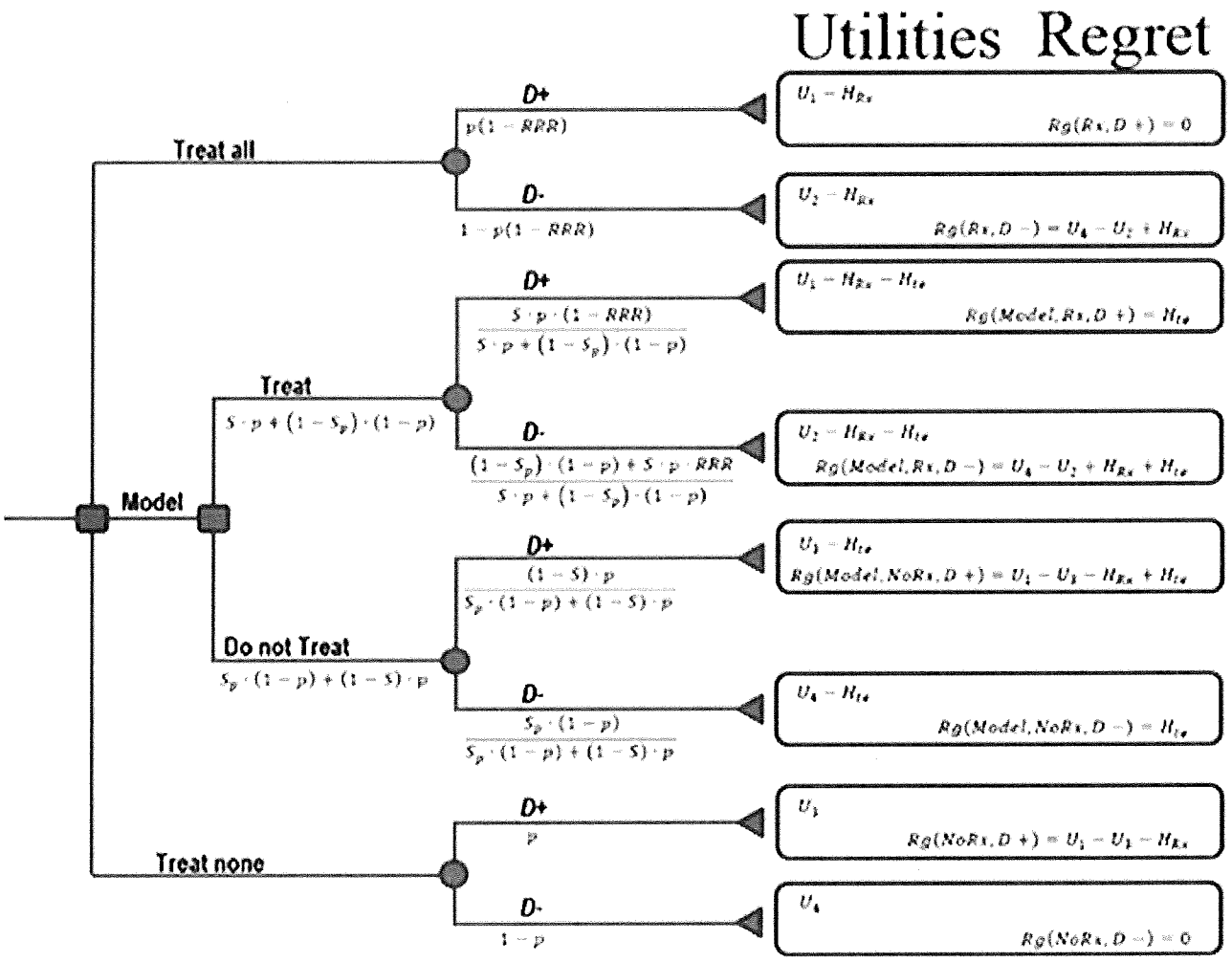
References

1. Centers for Disease Control and Prevention. "Leading Causes of Death". Available at: http://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm (last accessed 15 October 2015).
2. Cholesterol Treatment Trialists C (2012) The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. *The Lancet*, 380 (9841), 581–590.
3. Taylor, F., Huffman, M. D., Macedo, A. F., *et al.* (2013) Statins for the primary prevention of cardiovascular disease. *The Cochrane Database of Systematic Reviews*, (1), CD004816.
4. Stone, N. J., Robinson, J. G., Lichtenstein, A. H., *et al.* (2014) 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults – a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology*, 63 (25_Pt B), 2889–2934.
5. Jacobson, T. A., Ito, M. K., Maki, K. C., *et al.* (2015) National Lipid Association recommendations for patient-centered management of dyslipidemia: part 1 – executive summary. *Journal of Clinical Lipidology*, 8 (5), 473–488.
6. Ganda, O. P. (2015) Deciphering cholesterol treatment guidelines: a clinician's perspective. *JAMA: The Journal of the American Medical Association*, 313 (10), 1009–1010.
7. Eddy, D. M., Adler, J., Patterson, B., Lucas, D., Smith, K. A. & Morris, M. (2011) Individualized guidelines: the potential for increasing quality and reducing costs. *Annals of Internal Medicine*, 154 (9), 627–634.
8. D'Agostino, R. B. Sr, Vasan, R. S., Pencina, M. J., *et al.* (2008) General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*, 117 (6), 743–753.

9. Newman, C. B. & Tobert, J. A. (2015) Statin intolerance: reconciling clinical trials and clinical experience. *JAMA: The Journal of the American Medical Association*, 313 (10), 1011–1012.
10. Simon, H. A. (1955) A behavioral model of rational choice. *Quarterly Journal of Economics*, 69, 99–118.
11. Hume, D. (1748) Philosophical Essays Concerning Human Understanding. London: Millar.
12. Lichenstein, S. & Slovic, P. (2006) The Construction of Preference. New York: Cambridge University Press.
13. Djulbegovic, B., van den Ende, J., Hamm, R. M., Mayrhofer, T., Hozo, I. & Pauker, S. G. (2015) When is rational to order a diagnostic test, or prescribe treatment: the threshold model as an explanation of practice variation. *European Journal of Clinical Investigation*, 45 (5), 485–493.
14. Tsalatsanis, A., Hozo, I., Vickers, A. & Djulbegovic, B. (2010) A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making. *BMC Medical Informatics and Decision Making*, 10 (1), 51.
15. Pauker, S. G. & Kassirer, J. (1980) The threshold approach to clinical decision making. *The New England Journal of Medicine*, 302, 1109–1117.
16. Pauker, S. G. & Kassirer, J. P. (1975) Therapeutic decision making: a cost benefit analysis. *The New England Journal of Medicine*, 293, 229–234.
17. Djulbegovic, B. & Hozo, I. (2007) When should potentially false research findings be considered acceptable? *PLoS Medicine*, 4 (2), e26.
18. Hozo, I., Schell, M. J. & Djulbegovic, B. (2008) Decision-making when data and inferences are not conclusive: risk-benefit and acceptable regret approach. *Seminars in Hematology*, 45 (3), 150–159.
19. Zeelenberg, M. & Pieters, R. (2007) A theory of regret regulation 1.0. *Journal of Consumer Psychology*, 17, 3–18.
20. Hozo, I. & Djulbegovic, B. (2008) When is diagnostic testing inappropriate or irrational? Acceptable regret approach. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 28 (4), 540–553.
21. Hozo, I. & Djulbegovic, B. (2009) Clarification and corrections of acceptable regret model. *Medical Decision Making*, 29, 323–324.
22. Djulbegovic, B. & Paul, A. (2011) From efficacy to effectiveness in the face of uncertainty: indication creep and prevention creep. *JAMA: The Journal of the American Medical Association*, 305 (19), 2005–2006.
23. Tsalatsanis, A., Barnes, L. E., Hozo, I. & Djulbegovic, B. (2011) Extensions to Regret-based Decision Curve Analysis: an application to hospice referral for terminal patients. *BMC Medical Informatics and Decision Making*, 11, 77.
24. Djulbegovic, B., van den Ende, J., Hamm, R. M., *et al.* (2015) When is rational to order a diagnostic test, or prescribe treatment: the threshold model as an explanation of practice variation. *European Journal of Clinical Investigation*, 45 (5), 485–493.
25. Vickers, A. & Elkin, E. (2006) Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 26, 565–574.
26. Vickers, A. J., Cronin, A. M., Elkin, E. B. & Gonen, M. (2008) Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Informatics and Decision Making*, 8, 53.
27. Vickers, A. J. & Elkin, E. B. (2006) Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 26 (6), 565–574.
28. Pencina, M. J., Navar-Boggan, A. M., D'Agostino, R. B. Sr, *et al.* (2014) Application of new cholesterol guidelines to a population-based sample. *The New England Journal of Medicine*, 370 (15), 1422–1431.
29. Ioannidis, J. P. (2014) More than a billion people taking statins?: potential implications of the new cardiovascular guidelines. *JAMA: The Journal of the American Medical Association*, 311 (5), 463–464.
30. Zhang, H., Plutzky, J., Skentzos, S., *et al.* (2013) Discontinuation of statins in routine care settings: a cohort study. *Annals of Internal Medicine*, 158 (7), 526–534.
31. Zhang, H., Plutzky, J. & Turchin, A. (2013) Discontinuation of statins in routine care settings. *Annals of Internal Medicine*, 159 (1), 75–76.
32. Montori, V. M., Brito, J. P. & Ting, H. H. (2014) Patient-centered and practical application of new high cholesterol guidelines to prevent cardiovascular disease. *JAMA: The Journal of the American Medical Association*, 311 (5), 465–466.
33. Oshima Lee, E. & Emanuel, E. J. (2013) Shared decision making to improve care and reduce costs. *New England Journal of Medicine*, 368 (1), 6–8.
34. Cucchetti, A., Djulbegovic, B., Tsalatsanis, A., *et al.* (2014) When to perform hepatic resection for intermediate stage hepatocellular carcinoma. *Hepatology*, 61 (3), 905–914.
35. Hernandez, J. M., Tsalatsanis, A., Humphries, L. A., Miladinovic, B., Djulbegovic, B. & Velanovich, V. (2014) Defining optimum treatment of patients with pancreatic adenocarcinoma using regret-based decision curve analysis. *Annals of Surgery*, 259 (6), 1208–1214.
36. Zeelenberg, M. & Pieters, R. (2007) A theory of regret regulation 1.1. *Journal of Consumer Psychology*, 17, 29–35.
37. Stanovich, K. E. (2011) Rationality and the Reflective Mind. Oxford: Oxford University Press.
38. Djulbegovic, B., Elqayam, S., Reljic, T., *et al.* (2014) How do physicians decide to treat: an empirical evaluation of the threshold model. *BMC Medical Informatics and Decision Making*, 14 (1), 47.
39. Sorum, P. C., Mullet, E., Shim, J., Bonnin-Scaon, S., Chasseigne, G. & Cogneau, J. (2004) Avoidance of anticipated regret: the ordering of prostate-specific antigen tests. *Medical Decision Making*, 24 (2), 149–159.
40. Tsalatsanis, A., Hozo, I. & Djulbegovic, B. (2014) Empirical evaluation of regret and acceptable regret model. 36th Annual Meeting of Society of Medical Decision Making, Miami, October 19–22.
41. Sandberg, T. & Conner, M. (2008) Anticipated regret as an additional predictor in the theory of planned behaviour: a meta-analysis. *British Journal of Social Psychology*, 47 (4), 589–606.
42. Djulbegovic, M., Beckstead, J., Elqayam, S., *et al.* (2015) Thinking Styles and Regret in Physicians. *PLoS ONE*, 10 (8), e0134038.
43. Kahneman, D. (2003) Maps of bounded rationality: psychology for behavioral economics. *The American Economic Review*, 93, 1449–1475.
44. Stanovich, K. E. (2013) Why humans are (sometimes) less rational than other animals: cognitive complexity and the axioms of rational choice. *Thinking & Reasoning*, 19 (1), 1–26.
45. Djulbegovic, B., Hozo, I., Beckstead, J., Tsalatsanis, A. & Pauker, S. G. (2012) Dual processing model of medical decision-making. *BMC Medical Informatics and Decision Making*, 12 (1), 94.
46. Djulbegovic, B. & Hozo, I. (2010) Health Care Reform & Criteria for Rational Decision-making. Available at: http://smdm.org/uploads/general_files/SMDM_News_Spring_2010.pdf.
47. DeFilippis, A. P., Young, R., Carrubba, C. J., *et al.* (2015) An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohortcalibration and discrimination among CVD risk scores. *Annals of Internal Medicine*, 162 (4), 266–275.
48. Ridker, P. M. & Cook, N. R. (2015) Comparing cardiovascular risk prediction scorescomparing cardiovascular risk prediction scores. *Annals of Internal Medicine*, 162 (4), 313–314.

Appendix

Regret based decision curve analysis (DCA)



Definitions and assumptions (full description of the model given in [14,23])

Assume a patient comes in for a treatment and has probability of disease $P(D+) = P$. For the given value of threshold probability P_t, we will administer the treatment Rx, if $P(D+) \geq P$, and withhold treatment if $P(D+) < P_t$. However, for the patient undergoing the treatment, Rx, the probability that the patient actually has the disease is adjusted by RRR:$P(D+|R_x) = p\ (1 - RRR)$. Thus, if $RRR = 1$, disease is completely preventable; if $RRR = 0$, the treatment has no effect on the disease prevalence.

Net benefits of the treatment are defined as $B = U_1 - U_3$ [represents the difference in outcomes (disutilities) if the diseased patient were treated vs. not treated.]

Net harms/Risks of the treatment are defined as $H = U_4 - U_2$ [represents the difference in disutilities of not treated vs. treated in a patient without disease.]

Conditional probability versions of the positives and negatives rates for the given threshold without the treatment are given as:

True Positive Rate (Sensitivity, S): $TP = P(Treat|D+)$
$= P(p \geq P_t|D+)$

False Positive Rate: $FP = P(Treat|D-) = P(p \geq P_t|D-)$

True Negative Rate (Specificity, S_p): $TN = P(NoTreat|D-)$
$= P(p < P_t|D-)$

False Negative Rate: $FN = P(NoTreat|D+)$
$= P(p < P_t|D+)$

Note: $FN + TP = P(p < P_t|D+) + P(p \geq P_t|D+) = 1$ and
$FP + TN = P(p \geq P_t|D-) + P(p < P_t|D-) = 1$

The compound probabilities (intersections/percentages of total) are then given as the following for the given threshold P_t:

$\%FN = P(NoTreat \cap D+) = (1 - S) \cdot p = FN \cdot p$

$\%TN = P(NoTreat \cap D-) = S_p \cdot (1 - p) = TN \cdot (1 - p)$

$$\%TP = P(Treat \cap D+) = S \cdot p \cdot (1-RRR) = TP \cdot p \cdot (1-RRR)$$

$$\%FP = P(Treat \cap D-) = (1-S_p) \cdot (1-p) + S \cdot p \cdot RRR$$
$$= FP \cdot (1-p) + TP \cdot p \cdot RRR$$

Note: Under these assumptions, the total percentage of patients is divided into these four categories, $\%TP + \%FP + \%FN + \%TN = 1$.

The expected Regret (Erg) values are then derived from the tree diagram as:

$$ERg[Treat\ all] = p(1-RRR)(0) + (1-p(1-RRR))(U_4 - U_2 + H_{Rx})$$
$$= (1-p(1-RRR)) \cdot (H + H_{Rx})$$

$$ERg[Model] = H_{te} + (FP \cdot (1-p) + TP \cdot p \cdot RRR) \cdot (H + H_{Rx}) + p \cdot FN \cdot (B - H_{Rx})$$

$$ERg[Treat\ none] = p(U_1 - U_3 - H_{Rx}) + (1-p)(0) = p \cdot (B - H_{Rx})$$

The only difference between this model and the one we reported previously [14,23], is that in the original derivation we scaled the utilities by dividing each utility with the expression U_1–U_3, and replacing $\frac{H}{B} = \frac{U_4 - U_2}{U_1 - U_3} = \frac{P_t}{1-P_t}$. However, if we, instead of benefits, divide with the expression for harms $H = U_4$–U_2, and replace the ratio $\frac{B}{H} = \frac{U_1 - U_3}{U_4 - U_2} = \frac{1-P_t}{P_t}$ we obtain similar, but different expressions:

$$ERg^*[Treat\ all] = (1-p(1-RRR)) \cdot \left(1 + \frac{H_{Rx}}{H}\right) \quad (A1)$$

$$ERg^*[Treat\ none] = p \cdot \left(\frac{B}{H} - \frac{H_{Rx}}{H}\right) = p \cdot \left(\frac{1-P_t}{P_t} - \frac{H_{Rx}}{H}\right) \quad (A2)$$

$$ERg^*[Model] = \frac{H_{te}}{H} + (FP \cdot (1-p) + TP \cdot p \cdot RRR) \cdot \left(1 + \frac{H_{Rx}}{H}\right) + p \cdot FN \cdot \left(\frac{B}{H} - \frac{H_{Rx}}{H}\right)$$
$$= \frac{H_{te}}{H} + (FP \cdot (1-p) + TP \cdot p \cdot RRR) \cdot \left(1 + \frac{H_{Rx}}{H}\right) + p \cdot FN \cdot \left(\frac{1-P_t}{P_t} - \frac{H_{Rx}}{H}\right) \quad (A3)$$

$$NERD[Treat\ none, Treat\ all] = ERg^*[Treat\ none] - ERg^*[Treat\ all]$$
$$= p \cdot \left(\frac{1-P_t}{P_t} - \frac{H_{Rx}}{H}\right) - (1-p(1-RRR)) \cdot \left(1 + \frac{H_{Rx}}{H}\right) \quad (A4)$$

$$NERD[Treat\ none, model] = ERg^*[Treat\ none] - ERg^*[Model]$$
$$= TP \cdot p \cdot \left(\frac{1-P_t}{P_t} - \frac{H_{Rx}}{H}\right) - (FP \cdot (1-p) + TP \cdot p \cdot RRR) \cdot \left(1 + \frac{H_{Rx}}{H}\right) - \frac{H_{te}}{H} \quad (A5)$$

$$NERD[Treat\ all, model] = ERg^*[Treat\ all] - ERg^*[Model]$$
$$= (TN \cdot (1-p) + FN \cdot p \cdot RRR) \cdot \left(1 + \frac{H_{Rx}}{H}\right) - p \cdot FN \cdot \left(\frac{1-P_t}{P_t} - \frac{H_{Rx}}{H}\right) - \frac{H_{te}}{H} \quad (A6)$$

According to DCA theory, we say that strategy 1 is better than strategy 2 for an interval of threshold values $[a,b]$, if $NERD[Strategy1, Strategy2] < 0$ whenever $p_t \in [a,b]$.

Procedure

1 Suppose we have a group of patients for which the model assigns $P_i = p(D+)$ the probability of disease for the i^{th} patient, where $i = 1,2,3, \ldots, N$

2 Suppose that we actually know the reality for each patient, i.e. $d_i = 1$ (the i^{th} patient has the disease) or $d_i = 0$ (the i^{th} patient is disease free). Let $D_T = \sum_i d_i$ (the total number of patients with the disease) and estimate the variable $p = p(D+) = \frac{D_T}{N}$.

3 For each possible value of the threshold P_t, we will calculate the TP,FP,TN and FN as functions of the threshold P_t:

$$\text{True Positive}: TP = P(p \geq P_t | D+) = \frac{(\#\text{ of patients with } p_i \geq P_t \text{ and } d_i = 1)}{D_T}$$

$$\text{False Positive}: FP = P(p \geq P_t | D-) = \frac{(\#\text{ of patients with } p_i \geq P_t \text{ and } d_i = 0)}{N - D_T}$$

$$\text{True Negative}: TN = P(p < P_t | D-) = \frac{(\#\text{of patients with } p_i < P_t \text{ and } d_i = 0)}{N - D_T}$$

$$\text{False Negative}: FN = P(p < P_t | D+) = \frac{(\#\text{of patients with } p_i < P_t \text{ and } d_i = 1)}{D_T}$$

4 Calculate Expected Regret [formulas (1), (2) and (3)] or NERD [formulas (4), (5) and (6)] for each value of P_t.

Comparing two models

In a particular case when we have two models and want to compare the Net Expected Regret Difference, we simply repeat the procedure above for the threshold $P_t = x$ with particular (fixed) values of true and false positive and negative rates TP_x, FP_x, TN_x, and FN_x:

$$ERg^*[Model_x] = \frac{H_{te}}{H} + (FP_x \cdot (1-p) + TP_x \cdot p \cdot RRR) \cdot \left(1 + \frac{H_{Rx}}{H}\right) + p \cdot FN_x \cdot \left(\frac{1-x}{x} - \frac{H_{Rx}}{H}\right)$$

The Net Expected Regret Difference (NERD) between these two models is then

$$NERD[Model_x, Model] = ERg^*[Model_x] - ERg^*[Model]$$
$$= [(FP_x - FP)(1-p) + (TP_x - TP)\cdot p \cdot RRR]\left(1 + \frac{H_{Rx}}{H}\right)$$
$$+ p \cdot FN_x \cdot \frac{1-x}{x} - p \cdot FN \cdot \frac{1-P_t}{P_t} - (FN_x - FN)\cdot p \cdot \frac{H_{Rx}}{H} \quad \text{(A7)}$$

Overtreatment/Undertreatment formulas

Using the definitions of percentage count formulas, let's define the differential percentage counts (the differences between the percentage counts of two models) as:

$$Undertreatment: \%\Delta FN = \%FN_x - \%FN = FN_x \cdot p - FN \cdot p$$
$$= p \cdot (FN_x - FN) \quad \text{(A8)}$$

and

$$Overtreatment: \%\Delta FP$$
$$= (FP_x - FP)(1-p) + (TP_x - TP)\cdot p \cdot RRR \quad \text{(A9)}$$
$$= (FP_x - FP)(1-p) - (FN_x - FN)p \cdot RRR$$

Using these formulas, we can express the Net Expected Regret Difference using the overtreatment and undertreatment differences:

$$NERD[Model_x, Model]$$
$$= \%\Delta FP \cdot \left(1 + \frac{H_{Rx}}{H}\right) + \%FN_x \cdot \left(\frac{1-x}{x} - \frac{1-P_t}{P_t}\right) \quad \text{(A10)}$$
$$+ \%\Delta FN \cdot \left(\frac{1-P_t}{P_t} - \frac{H_{Rx}}{H}\right)$$

In general – overtreatment can be calculated as the difference in False Positives between two different strategies/alternatives, like in formulas (8) and (9) above. Undertreatment is, similarly, the difference in False Negatives. Therefore, we have the following False Positive and False Negative rates for various strategies: Treat None ($FP = 0$ and $FN = 1$), Treat All ($FP = 1$ and $FN = 0$), Threshold Model ($FP = FP(P_t)$ and $FN = FN(P_t)$) and Guideline Model ($FP_x = FP(x)$ and $FN_x = FN(x)$).

The differences are then in general given as:

$$Undertreatment: \%\Delta FN(Strategy_1, Strategy_2) = p \cdot (FN_1 - FN_2)$$
$$Overtreatment: \%\Delta FP(Strategy_1, Strategy_2)$$
$$= (FP_1 - FP_2)(1-p) - (FN_1 - FN_2)\cdot p \cdot RRR$$

$$Undertreatment: \%\Delta FN(Treat\ All, Treat\ None) = p \cdot (0-1)$$
$$= -p$$
$$Overtreatment: \%\Delta FP(Treat\ All, Treat\ None)$$
$$= (1-0)(1-p) - (0-1)\cdot p \cdot RRR = (1-p) + p \cdot RRR$$

$$Undertreatment: \%\Delta FN(Model, Treat\ None) = p \cdot (FN - 1)$$
$$Overtreatment: \%\Delta FP(Model, Treat\ None)$$
$$= (FP - 0)(1-p) - (FN - 1)\cdot p \cdot RRR$$
$$= FP \cdot (1-p) + (1 - FN)\cdot p \cdot RRR$$

$$Undertreatment: \%\Delta FN(Guideline, Treat\ None) = p \cdot (FN_x - 1)$$
$$Overtreatment: \%\Delta FP(Guideline, Treat\ None)$$
$$= (FP_x - 0)(1-p) - (FN_x - 1)\cdot p \cdot RRR$$
$$= FP_x(1-p) + (1 - FN_x)\cdot p \cdot RRR$$

$$Undertreatment: \%\Delta FN(Model, Treat\ All) = p \cdot (FN - 0)$$
$$= p \cdot FN$$
$$Overtreatment: \%\Delta FP(Model, Treat\ All)$$
$$= (FP - 1)(1-p) - (FN - 0)\cdot p \cdot RRR$$
$$= (FP - 1)(1-p) - FN \cdot p \cdot RRR$$

$$Undertreatment: \%\Delta FN(Guideline, Treat\ All) = p \cdot (FN_x - 0)$$
$$= p \cdot FN_x$$
$$Overtreatment: \%\Delta FP(Guiedline, Treat\ All)$$
$$= (FP_x - 1)(1-p) - (FN_x - 0)\cdot p \cdot RRR$$
$$= (FP_x - 1)(1-p) - FN_x \cdot p \cdot RRR$$

$$Undertreatment: \%\Delta FN(Model, Guideline) = p \cdot (FN - FN_x)$$
$$Overtreatment: \%\Delta FP(Model, Guideline)$$
$$= (FP - FP_x)(1-p) - (FN - FN_x)\cdot p \cdot RRR$$

The "*" means this is SCALED expected regret. Scaled by H=U4-U2.

Djulbegovic et al. BMC Medical Informatics and Decision Making 2014, 14:47
http://www.biomedcentral.com/1472-6947/14/47

BMC
Medical Informatics & Decision Making

RESEARCH ARTICLE Open Access

How do physicians decide to treat: an empirical evaluation of the threshold model

Benjamin Djulbegovic[1,2,3,8*], Shira Elqayam[4], Tea Reljic[1], Iztok Hozo[5], Branko Miladinovic[1], Athanasios Tsalatsanis[1], Ambuj Kumar[1,2], Jason Beckstead[6], Stephanie Taylor[1] and Janice Cannon-Bowers[1,7]

Abstract

Background: According to the threshold model, when faced with a decision under diagnostic uncertainty, physicians should administer treatment if the probability of disease is above a specified threshold and withhold treatment otherwise. The objectives of the present study are to a) evaluate if physicians act according to a threshold model, b) examine which of the existing threshold models [expected utility theory model (EUT), regret-based threshold model, or dual-processing theory] explains the physicians' decision-making best.

Methods: A survey employing realistic clinical treatment vignettes for patients with pulmonary embolism and acute myeloid leukemia was administered to forty-one practicing physicians across different medical specialties. Participants were randomly assigned to the order of presentation of the case vignettes and re-randomized to the order of "high" versus "low" threshold case. The main outcome measure was the proportion of physicians who would or would not prescribe treatment in relation to perceived changes in threshold probability.

Results: Fewer physicians choose to treat as the benefit/harms ratio decreased (i.e. the threshold increased) and more physicians administered treatment as the benefit/harms ratio increased (and the threshold decreased). When compared to the actual treatment recommendations, we found that the regret model was marginally superior to the EUT model [Odds ratio (OR) = 1.49; 95% confidence interval (CI) 1.00 to 2.23; p = 0.056]. The dual-processing model was statistically significantly superior to both EUT model [OR = 1.75, 95% CI 1.67 to 4.08; p < 0.001] and regret model [OR = 2.61, 95% CI 1.11 to 2.77; p = 0.018].

Conclusions: We provide the first empirical evidence that physicians' decision-making can be explained by the threshold model. Of the threshold models tested, the dual-processing theory of decision-making provides the best explanation for the observed empirical results.

Keywords: Medical decision-making, Threshold model, Dual-processing theory, Regret, Expected utility theory

Background

Medical decision-making is often performed under conditions of diagnostic uncertainty; that is, physicians frequently need to decide whether to give treatment to a patient who may or may not have a disease. Clinical practice is full of these examples. For instance, if the physician treating a patient with a sore throat estimates that the probability of streptococcal infection is sufficiently high, she may decide to treat – assuming that the benefits of administering antibiotic outweigh its potential harms. Thus, to make appropriate therapeutic decision when a diagnosis is uncertain, the clinician has to: 1) ascertain the probability of a patient having the disease, and 2) decide whether the potential treatment benefits will outweigh its harms.

In everyday clinical practice, the assessment of the likelihood of disease and balance of treatment's benefits and harms is often done intuitively, but this decision-making process can be formalized under the "threshold model" [1,2]. According to the threshold model, when faced with uncertainty about whether to treat a patient who may or may not have a disease, there must exist some probability at which a physician is indifferent

* Correspondence: bdjulbeg@health.usf.edu
[1]Department of Internal Medicine, Division of Evidence-based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL, USA
[2]Department of Health Outcomes and Behavior, Moffitt Cancer Center & Research Institute, Tampa, FL, USA
Full list of author information is available at the end of the article

between administering versus not administering treatment; this is known as threshold probability [1,2]. Physicians would choose to treat when the probability of disease is above the threshold probability and would choose to withhold treatment otherwise [1,2]. The threshold model stipulates that as the therapeutic benefit/harms ratio increases, the threshold probability at which treatment is justified is lowered. Conversely, if the treatment's benefit/harms ratio decreases, the required threshold for therapeutic action will be higher. To date, three types of threshold models have been described: 1) the original model, based on the expected utility theory (EUT) framework (T_{EUT}) [1,2]; 2) the regret-based threshold model (T_{RG}) [3-5]; and 3) the threshold model based on the dual-processing theory of decision-making (T_{DP}) [6].

The T_{EUT} model is derived from the principles of decision theory, which hold that a decision-maker should select the option with the highest expected utility to maximize achievement of valued outcomes. The T_{RG} model is based on expected regret theory, which holds that the preferred course of action is based on the least amount of regret associated with a possibly wrong decision. The T_{DP} model is based on dual processing theories, which postulate that our cognition is governed by so called type 1 or 2 processes [7-15]. Type 1 processes are intuitive, automatic, fast, narrative, experiential and affect-based; type 2 processes are analytical, slow, verbal, and deliberative supporting formal logical and probabilistic analyses [7-16].

Despite the widespread popularity, none of the threshold models (T_{EUT}, T_{RG}, T_{DP}) have been submitted to empirical evaluation to test their descriptive accuracy. The purpose of our study was to assess whether physicians act according to a threshold model, and if they do, to determine which model best explains their decision-making. Knowing if physicians operate under a threshold model and which model best describes physicians' decisions is very important for medical education as it can help identify the most salient features of medical decision-making. This, in turn can be used for didactic purposes towards better practice of clinical decision-making. In addition, understanding the decision-making processes can help explain patterns observed in the contemporary clinical practice such as treatment overuse and underuse.

Methods

Participants and setting

Physicians from the University of South Florida and Evidence-based Medicine Discussion Group were recruited for the study via email invitation to participate in a web-based survey. E-mail invitations were sent via institutional listserv followed by a weekly reminder. No incentives were offered for participation in the study. The only inclusion criteria were that participants were practicing physicians, regardless of the field of medicine, actively involved in therapeutic decision-making on a daily basis. The survey was closed after the target sample was reached. The study was approved by the USF IRB (No. Pro9047).

Design and materials

All theories of decision-making agree that choices are functions of benefits (gains) and harms (losses). Therefore, we constructed the case vignettes to allow easy discernment of benefits and harms for serious, life-threatening outcomes. The aim was to compel our study participants to rely on the estimates of benefits and harms, in particular on the benefit/harm (B/H) ratio. To minimize "framing effect" [17], we chose presentation and wording that is commonly used in the literature and medical communication and with which most physicians are familiar.

Threshold models

Our case vignettes refer to a clinical situation when a decision about treatment has to be made but a physician is uncertain whether the patient has a given condition and no further diagnostic tests are available to her/him to reduce the diagnostic or prognostic uncertainty. We now provide a brief outline of all 3 models:

1) Expected utility threshold model

Although often considered gold standard of rationality, violation of decision-making by EUT is well documented in literature [5,18-21]. However, one issue is rarely directly addressed: do people violate precepts of EUT because of errors due to brain processing limitations, or because EUT does not reflect the optimal decision-making perspective of the decision-maker. For example, few people can accurately multiply 3.4578*4,678; that does not, however, mean they reject (normatively) the correct answer once they perform the calculation with help of a calculator. Most people simply correct their error and accept the answer obtained after punching the numbers into a calculator. We, therefore, asked the following question: will people behave according to EUT after they are told what they should (normatively) do? Or, will they violate the rules of EUT even after they are told what is the theoretical best course of action? For this purpose, we included a number of prescriptive statements in our case vignettes based on the EUT normative calculations.

The EUT threshold was calculated as:

$$T_{EUT} = 1/\left(1 + \frac{B_2}{H_2}\right) \quad (1)$$

where benefits/harms (B_2/H_2) refer to the objective data obtained from the literature. Thus, if $B_2/H_2 = 9$, the probability above which we should give treatment is only

10%. [The EUT model relies on type 2 processes. Hence, we used the subscript 2 in equation 1].

2) Regret threshold model

Many clinical decisions are driven by regret where a decision-maker (a doctor or a patient) seeks to minimize regret associated with a potentially wrong decision [3-5]. In general, in a clinical situation similar to the one considered here, a decision maker deals with two types of regret: failure to provide benefit (regret of omission) versus administering unnecessary and potentially harmful treatment (regret of commission) [3-5]. Given that in medical decision-making most decisions cannot be reversed (e.g., once surgery has occurred, its effects cannot be reversed), the T_{RG} model is based on anticipatory regret only [3-5]) (as opposed to retrospective regret or post-decision justification regret [22,23]). Anticipation of regret leads to more vigilant decision making, satisfying most of the criteria of high-quality decisions [8,24]. To estimate regret of omission versus commission, as alluded above, we employed the regret-based Dual Visual Analog Scale (DVAS) [25] (see Figure 1 and Additional file 1 for further details on actual regret elicitation). Regret threshold was calculated by employing the following formula:

$$T_{REG} = 1/\left(1+\frac{B_1}{H_1}\right) \quad (2)$$

where B_1/H_1 is failure to benefit/unnecessary harms. Note the regret threshold model is, psychologically, a type 1 only model, which relies on holistic assessment of benefits and harms (hence, we used subscript 1 in the equation). That is, the model predicts that the responses will be determined by regret, which is an affective (and hence type 1) response.

3) Dual-processing threshold model

In recent years, it has become evident that decision-making theories which assume a single system of reasoning are not sufficient to explain human decision-making [8,9,26-28]. Instead, as introduced above, it is increasingly accepted that cognitive processes are governed by both type 1 and type 2 processes [8,9,26-28]. We recently developed a threshold model based on dual processing theory (T_{DP}), which takes into account analytical type 2 functioning based on rational calculus of EUT as well as type 1 mechanisms driven both by emotion (regret) and other type 1 processes [6].

The decision to administer treatment according to type 2 processing depends on the EUT threshold calculated as shown in equation 1. The extent of type 1 processes (i.e., the extent to which type 1 processes are not suppressed by or compete with type 2 processes) in the decision-making is given by parameter γ [0 to 1]; if $\gamma = 0$, then decision-making adheres to EUT. Conversely, if $\gamma = 1$, then type 1 processes dominate decision-making. For any $0 \le \gamma \le 1$, decision-making is a combination of both processes. The formula for calculation of the T_{DP} is given by:

$$T_{DP} = (T_{EUT})\left[1+\frac{\gamma}{2(1-\gamma)}\left(\frac{H_1}{H_2}\right)\left(1-\frac{B_1}{H_1}\right)\right] \quad (3)$$

As explained, B_1 and H_1 are elicited from the participants (Figure 1) while T_{EUT} is calculated based on the best evidence from the literature, B_2 and H_2. Because γ represents the extent of activation of type 1 processes, this can be conceptualized as relative distance between analytically derived T_{EUT} and regret-based, T_{REG}. Thus, we calculated γ in the following way (keeping the value between 0 and 1):

$$\gamma = \begin{cases} \frac{T_{EUT}-T_{RG}}{T_{EUT}}, if \frac{T_{EUT}-T_{RG}}{T_{EUT}} < 1 \\ 1, otherwise \end{cases} \quad (4)$$

Therefore, γ is equal to $\frac{T_{EUT}-T_{RG}}{T_{EUT}}$, if $\frac{T_{EUT}-T_{RG}}{T_{EUT}} < 1$. If $\frac{T_{EUT}-T_{RG}}{T_{EUT}} \ge 1$, then γ is equal to 1. Estimates for γ are provided in Additional file 2, Table S1.

Note that there are many dual-processing theories [29] and the model presented here represents a specific dual-processing model that is applicable to single-point clinical decisions [6].

A survey to test the threshold models

We devised two clinical scenarios - one for a familiar condition and a second which required specialized knowledge. Scenario 1 was about treatment of pulmonary embolism (PE), which should be familiar to the vast majority of physicians. Scenario 2 was about treatment of acute myeloid leukemia (AML), with which only a minority of physicians have experience (see Additional file 2 for the survey/concrete examples).

To examine dual processing aspects, we used a variation of the two-response paradigm in which initial responses are considered to represent mostly type 1 processes, and later responses are considered to represent the added influence of type 2 processes. We, therefore, included more detailed information between the first and the second response.

To capture this initial (type 1) response, we first asked all participants to provide their best assessment on benefits/harms for treatment of PE and AML, respectively. That is, the first question was devoid of any case-specific contextual details. This response to benefits (B) and harms (H) due to over-learned processes (see below and Discussion) is postulated to be automatic ($_{aut}$), and we label them here as B_{aut} and H_{aut}.

The B_{aut} over H_{aut} is stipulated to serve as an "anchor" but is expected to be further modified by the contextual

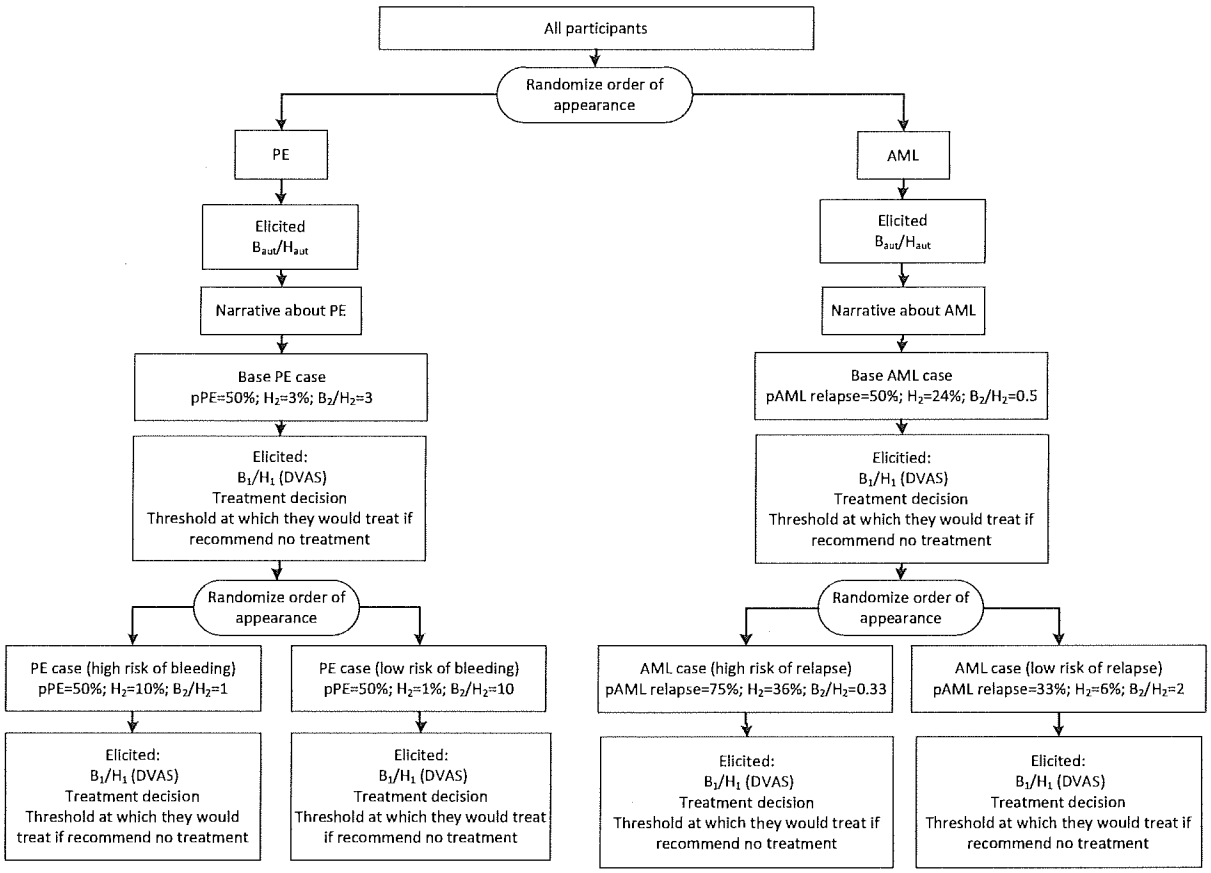


Figure 1 A schema of the experimental design. Note that design was entirely within participants and that all participants answered all question (but in different order, according to the randomization sequence). Abbreviations: PE, Pulmonary embolism; AML, Acute Myeloid Leukemia; B_{aut}/H_{aut}, automatic benefit to harm ratio; pPE, probability of PE, H2, harms associated with treatment provided; B_2/H_2, benefit to harm ratio provided in the case; B_1/H_1, benefit to harm ratio elicited form participants using DVAS; Dvas, dual visual analog scale; pAML, probability of AML relapse. Note: All participants completed all vignettes. Only the order of presentation of vignettes was randomized where indicated.

details of each case presentation as affected by the various type 1 and type 2 processes. By eliciting the anchor value, our attempt was to ensure elicitation of the subsequent responses related to B_1 and H_1 estimates within clinically realistic range. Note, however, we only need to elicit B_1 and H_1 values to perform the actual calculations; elicitation of B_{aut} and H_{aut} only serve to conduct the experimental procedure according to our theoretical framework.

We note that type 1 processes are determined by a number of factors, including: (a) affect, (b) evolutionary hard-wired processes, responsible for automatic responses to potential danger, (c) over-learned processes based on type 2 mechanisms that have been relegated to type 1 responses (such as the effect of intensive training resulting in the use of heuristics), and (d) the effects of tacit learning [11]. All these factors were taken into account in construction of the vignettes in the following way: medical education and exams typically consist of case vignettes, which after many hours of training become internalized and represent the basis for acquiring expertise and actual practice of medicine. The vignettes, therefore, were constructed to be as realistic as possible in order to represent actual patients with additional context-specific details. Thus, the response to the case integrates automatic type 1 processes to capture both the effect of intensive training (which relies on the use of heuristics) and affect (regret) to possible acts of omission or commission associated with potentially wrong treatment. The latter was measured using DVAS for assessment of regret in holistic fashion [25] (See also Additional file 1). That is, the regret-related consequences had encompassed all possible harms and benefits envisioned by the respondents. Therefore, we label actually elicited benefits and harms as B_1 and H_1.

To activate type 2 deliberations and analytic processes, we provided additional objective data on the management of PE and AML based on the best available evidence in the literature. This was given both in terms

Table 1 Participant demographics and experience

Variable	Number of participants (%)
Overall	41 (100)
Gender	
Male	28 (68)
Female	13 (32)
Age	
Median (Range)	41 (26 to 66)
Area of specialization	
Anesthesiology	2 (5)
Dermatology	1 (2)
Emergency Medicine	1 (2)
Family Medicine	10 (24)
Hematology and Oncology	14 (34)
Internal Medicine	5 (12)
Obstetrics and Gynecology	2 (5)
Otolaryngology	1 (2)
Pediatrics	1 (2)
Urology	2 (5)
Other*	2 (5)
Level of experience	
Resident	10 (24)
Fellow	8 (20)
Attending	23 (56)
Experience treating patients for PE (N = 41)	
None	3 (7)
Fewer than 5 patients	11 (27)
Between 5 and 10 patients	4 (10)
Between 11 and 20 patients	7 (17)
More than 20 patients	16 (39)
PE vignettes similar to experience (N = 38)	
Yes	30 (79)
No	8 (21)
Experience treating patients for AML (N = 41)	
None	25 (61)
Fewer than 5 patients	4 (10)
Between 5 and 10 patients	1 (2)
Between 11 and 20 patients	4 (10)
More than 20 patients	7 (17)
AML vignettes similar to experience (N = 16)	
Yes	14 (88)
No	2 (12)
Understand formal principles of decision analysis (N = 41)	
Yes	29 (71)
No	12 (29)

*One public health and one preparing for residency in internal medicine.

of general narrative description of treatment for PE and AML and specific prescriptive statements that "treatment is justified when probability of disease (PE or AML) is sufficiently high for given benefits and harms". We label the objective benefits and harms as B_2 and H_2, respectively.

To keep the scenarios as realistic as possible, benefit and harms parameters were tailored to the case descriptions (PE, AML). Benefits and harms were given for each case (6 vignettes in total). Three vignettes included description of PE and three described AML cases. The three vignettes represented the base-case (intermediate benefits/harms ratio), high-risk (with low benefit/harms ratio resulting in higher threshold in comparison with the base-case), and low-risk (high benefit/harms ratio resulting in lower threshold in comparison with the base-case). In the vignettes, we also provided data on probability of disease (PE or AML relapse, respectively). In addition, when asked "would you give treatment to this patient" in the instruction prior to presenting the first (base-case) vignette, we included a normative statement that "treatment should be given if probability of disease exceeds probability X" where X was derived using B_2/H_2 data and referred to the probability of PE and AML, respectively. In PE vignettes, in addition to providing assessment of probability of disease in a base-case vignette, we also included data on the probability of PE in high- and low-risk vignettes (we kept probability of PE in all scenarios at 50%). The intent was to enable type 2 functioning to the maximum possible extent, and to ensure that the observed results are not ascribed to simple error in calculations but rather reflect activation of systematic cognitive processes (see also below). In case of AML, we provided sufficient details from which a physician familiar with treatment of AML could easily deduce high or low probability of relapse (but without including explicit quantitative statements about probability of AML relapse). The intent here was to simulate actual practice where experts typically talk about "high" or "low" risk for relapse, but rarely quantify it. In both cases, we expected to observe the physicians' behavior according to a threshold model.

Finally, to control for the order of presentation, we randomly presented PE versus AML vignettes. We further randomized the order of presentation to low versus high "threshold" descriptions, and the DVAS anchor used to elicit regret (i.e. we randomized a default slider position at 0% vs. 100%). Thus, all participants were presented all questions related to all vignettes, but the ordering of questions was randomized within the individual participants.

In summary, the manipulated factors were: response stage (initial/final), scenario familiarity (pulmonary embolism/acute myeloid leukemia), and level of threshold ("risk") according to EUT (high/low B_2/H_2 ratio), all manipulated within participants.

Figure 1 shows details of the experimental design.

Statistical analysis

We planned to recruit 40 participants, which is a customary sample size for cognitive psychology experiments. To test our main hypothesis, we postulated the following: if the threshold concept operates, then fewer physicians will give treatment as the threshold probability increases; this is because the physicians will require higher diagnostic certainty to prescribe treatments when threshold level is high. Conversely, as the threshold drops, lower diagnostic certainty is required, and more physicians will prescribe treatment. To assess whether our predictions will bear out, we compared responses to the base-case vignettes with those in which the threshold was higher ("high-risk", low B_2/H_2) or lower ("low-risk", high B_2/H_2) in relation to the base-case scenario. Thus, the main outcome in our study was comparison of a proportion of the physicians who will or will not prescribe treatment in relation to perceived change in the EUT threshold probability. To assess for the difference in responses between base-case and high-risk (low B_2/H_2, high threshold) and base-case and low-risk (high B_2/H_2, low threshold) scenarios we employed McNemar's test because of the paired nature of our data [30].

Our secondary outcomes consisted of deriving three thresholds, one for each model (i.e., T_{EUT}, T_{RG} and T_{DP}) with respect to the given probability of diagnosis of PE and AML relapse, respectively. We postulated that the actual threshold would be lower than the estimated probability of disease for physicians who decided to treat. On the other hand, for physicians who decided not to treat, the threshold will be higher than the estimated probability of disease. We computed the threshold for each participant and assessed whether their decisions to treat or not were in agreement with the particular threshold model. To explain which threshold model can best explain our main results, we assessed the difference in agreement between all three threshold models. Agreement was established if the probability of PE or AML was greater than or equal to threshold and the participant decided to treat or if the probability of PE or AML was less than threshold and the participant decided not to treat. A two-level logit mixed-model was applied which allowed us to account for the correlated multiple responses within each participant for each of the six vignettes. The model was fit using the command meqrlogit in STATA [31].

Results

A total of 41 consecutively enrolled physicians participated in the web-based survey. Two out of 41 participants were not practicing physicians (1 was a public health professional, and 1 was preparing for residency in internal medicine). Data from these two participants were included in the report as there were no significant differences in the findings when they were removed from the analysis. To ensure that we enrolled a sufficient number of physicians with experience in treating AML, an invitation to participate was first sent to hematology and oncology fellows and the faculty at the USF. After receiving 10 responses, we sent invitations for the survey to all other types of specialties. Details on the demographics of participants and other characteristics are summarized in Table 1. Thirty-eight of the 41 participants (93%) had experience treating PE, while 16 (39%) of physicians had experience with treatment of patients with AML. Both PE and AML vignettes were judged by majority of physicians (79% and 88%, respectively) as realistic examples of real-life clinical situations. Twenty-nine (71%) participants stated that they are familiar with the formal principles of decision analysis (which is based on EUT).

Table 2 shows the results of main analysis. The results are consistent with our main hypothesis: fewer physicians treat as the benefit/harms ratio decreased (i.e. threshold increased) whereas more physicians administered treatment as the benefit/harms ratio went up (and the threshold decreased). A significantly lower proportion of physicians favored treatment in the "high threshold" (high-risk) case compared to the base-case both for PE and AML case vignettes ($p < 0.0001$). Similarly, a significantly higher proportion of physicians favored treatment in the "low threshold" (low-risk) case compared to the base-case ($p < 0.0001$) in the AML vignette. However, there were no statistically significant differences in responses between the base-case and "low threshold" case for PE. The reason for this is that, surprisingly, we detected ceiling effects in the PE case: all physicians stated that they would treat the patient in the vignette with high benefit/harm ratio ("low-risk", "low threshold" vignette) while only one physician would not treat the patient in the base-case vignette. Nevertheless, qualitatively the results went in the same direction providing overall support for our hypotheses. In addition, the results were robust to the sensitivity analyses according to the years of experience, areas of expertise, familiarities with the clinical situation, experience with decision analysis, or order of randomization (see sensitivity analysis in Table two in Additional file 1). Thus, the findings indicate that, relative to base rates, the probability of treatment decreased in the "high threshold" ("high-risk", low benefit/harm ratio) vignettes, and increased in the "low threshold" ("low-risk", high benefit/harm ratio) vignettes (except for PE where treatment probability was at ceiling in the base-case and could not increase any further).

The results show that the threshold concept is likely to be operating in clinical practice but does not clarify which threshold model is valid (Table 2). Table 3 shows the threshold value results according to all three threshold models tested (Additional file 2). When compared to the actual

Table 2 Decision to administer treatment (N = 41)

	Pulmonary Embolism					Acute Myeloid Leukemia				
Case	Treat (%)		No treat (%)		p-value	Treat (%)		No treat (%)		p-value
Base case	40	(98)	1	(2)		27	(66)	14	(34)	
High threshold ("risk") case	16	(39)	25	(61)	<0.0001	8	(20)	33	(80)	<0.0001
Low ("risk") threshold case	41	(100)	0	(0)	1	36	(88)	5	(12)	0.012

treatment recommendations in a pooled mixed model analysis, we found that the regret model was marginally statistically superior to the EUT model [Odds ratio (OR) = 1.49; 95% confidence interval (CI) 1.00 to 2.23; p = 0.06]. The dual-processing model was statistically significantly superior to both the EUT model [OR = 1.75, 95% CI 1.67 to 4.08; p < 0.001] and regret model [OR = 2.61, 95% CI 1.11 to 2.77; p = 0.018]. Figure 2 shows predicted probability of the agreeing with threshold for each model. Thus, the dual-processing threshold model appears to most consistently agree with the observed data.

Discussion

In this paper, we provide empirical evidence that physicians appear to make their decisions according to the threshold model. A few empirical studies evaluated if physicians make decisions according to the threshold model [18,19] but none consider putting their results within a specific theoretical framework such as regret or dual processing theories. In this paper, we evaluated three types of threshold models published in the literature so far: 1) EUT [2], 2) regret [3,4], and 3) dual-processing model [6].

Regardless which threshold model can explain physicians' treatment decisions best, our finding that the threshold model appears to underpin typical clinical decision-making has practical implications for the practice of medicine and medical education. For example, it is estimated that between 30-50% of health care represents waste, mostly due to over-treatment [32]. Furthermore, approximately 80% of all health care expenditures are attributed to physicians' decisions [33]. If physicians' do act according to the threshold model, this would mean that every time they perceive that benefits of a treatment substantially outweigh its harms, we can expect that the treatment threshold will predictably drop. The lower the threshold, the lower is the diagnostic certainty required to justify treatment, thereby leading more physicians to prescribe treatment [5,20,21,34]. While this behavior may be rational, it, in turn, will lead to increase in over-treatment [5]. For example, in the baseline case of PE, almost all physicians (98%) would commit to treatment even though probability of PE was only 50%; that is, almost half of patients without PE would be treated unnecessarily. Conversely, the requirement for higher diagnostic certainty may lead to under-treatment. For example, in the high threshold case, only 39% of physicians would give treatment, even though the probability of PE was 50% (Table 2). Thus, depending on the clinical circumstances, both under- and over-treatment do occur in

Table 3 Physicians whose decision to administer treatment was in agreement with specific threshold (N = 41)

	Pulmonary Embolism						Acute Myeloid Leukemia					
	Agree	(%)	Disagree	(%)	EUT versus regret p-value	EUT or regret versus dual p-value	Agree	(%)	Disagree	(%)	EUT versus regret p-value	EUT or regret versus dual p-value
Base case												
EUT	40	(98)	1	(2)		1	27	(66)	14	(34)		0.096
Regret	38	(93)	3	(7)	0.625	0.625	33	(80)	8	(20)	0.146	0.727
Dual	40	(98)	1	(2)			35	(85)	6	(15)		
High risk case												
EUT	16	(39)	25	(61)		0.004	8	(20)	33	(80)		<0.001
Regret	31	(76)	10	(24)	0.003	1	25	(61)	16	(39)	<0.001	<0.001
Dual	30	(73)	11	(27)			40	(98)	1	(2)		
Low risk case												
EUT	41	(100)	0	(0)		<0.001	36	(88)	5	(12)		0.453
Regret	37	(90)	4	(10)	0.125	0.118	23	(56)	18	(14)	0.011	0.021
Dual	30	(73)	11	(27)			33	(80)	8	(44)		

Note: Agreement was established if the probability of PE or AML was greater than or equal to threshold and the participant decided to treat or the probability of PE or AML was less than threshold and the participant decided not to treat.

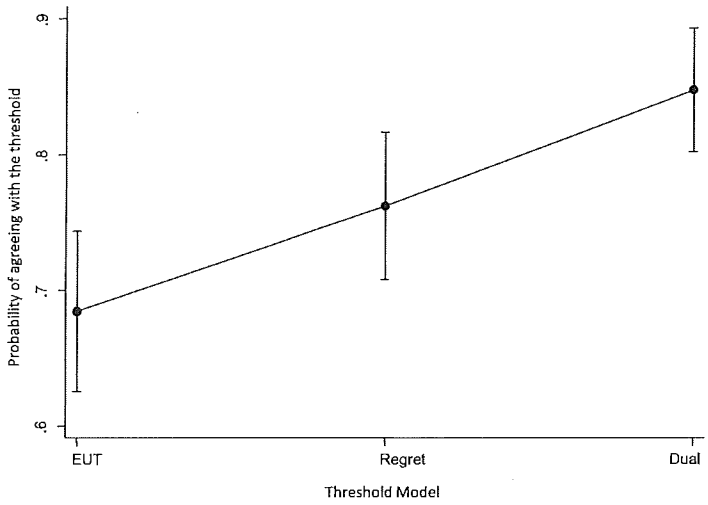


Figure 2 The predicted probability of the agreeing with threshold for each model. Dual processing model seems to fit the data best.

current medical practice and can be explained by the threshold model [4-6]. In general, however, over-treatment dominates the current medical practice in the US [33,35].

Overall, the EUT model predicted the observations with less accuracy compared to regret and dual-processing based models. Although finding that people violate expected utility theory is not new [8,20,21,36-38] it is, however, most interesting that many physicians did not act according to the EUT despite being given prescriptive advice indicating that it may be the most rational approach and regardless of the fact that the majority of them have been exposed to formal principles of decision analysis. The participants satisfied all the criteria for normative response: they had sufficient cognitive ability, high motivation, and appropriate 'mindware' i.e., cognitive tools to apply to the task [11], yet they failed to do so. We are not aware of any literature where this has been documented; in fact one lingering question related to the literature about violation of EUT relates to the issue whether the results can be explained by simple computational processing errors in the way people manipulate data on outcomes and probabilities. Our findings show that it is not simple processing errors that led to rejection of EUT. Rather, the results point to the fundamental findings that physicians, like other people [39], do not appear to follow prescriptive EUT as the optimal decision-making framework for medical decision-making. These observations have implications for practice of medicine as influential organizations charged to make clinical recommendations such as the United States Preventive Services Task Force (USPSTF) have increasingly used modeling based on EUT to issue clinical recommendations [40]. The fact that physicians may fail to follow EUT as a basis for decision-making may explain, for example, the vociferous debate that accompanied publication of the USPSTF guidelines on screening mammography [41].

We expected that much of the physicians' actions are driven by automatic type 1 processes further modified by the contextual details of a given clinical situation. This is the consequence of the way medical education is structured, as the overlearned processes from thousands of hours of training eventually become one's second nature that serve as the basis for quick, automatic decisions. We found that regret-based B_1/H_1 did differ from B_{aut}/H_{aut} ratios across presented scenarios (Table 4). This, as stipulated in the Methods, indicates that the contextual characteristics of the cases presented in the vignettes

Table 4 Benefit versus harm ratio based on type 1 response*

Variable		n	Mean	Min	Median	Max
PE	B_{aut}/H_{aut}	40	4.33	.6	3.00	25.00
Base case	B_1/H_1	40	6.28	0.75	3.18	49.50
Low risk	B_1/H_1	39	12.46	0.66	5.26	100.00
High risk	B_1/H_1	41	1.76	0.05	0.98	18.80
AML	B_{aut}/H_{aut}	41	2.29	0.43	2.00	10.00
Base case	B_1/H_1	41	1.55	0.00	1.00	7.07
Low risk	B_1/H_1	39	4.39	0.00	1.94	22.50
High risk	B_1/H_1	40	0.70	0.00	0.50	3.00

Abbreviations: B_{aut}/H_{aut} assessment of benefit/harms ratio based on automatic, quick response, B_1/H_1-type 1 response driven by regret, *PE* pulmonary embolism, *AML* acute myeloid leukemia, *low "risk"* low threshold, *high "risk"* high threshold clinical decisions. [*Note that type 2 responses that relied on single values, fixed B_2/H_2 ratios precluding direct statistical comparisons with B_{aut}/H_{aut}. However, the values of B_2/H_2 differed considerably from B_{aut}/H_{aut} (from 1 to 10 in PE case, and 2 to 0.33 in AML case) consistent with a notion that the B_{aut}/H_{aut} estimates did not solely drive the decision-making (see Discussion)].

triggered other cognitive mechanisms both along the type 1 (e.g., regret) and type 2 processes.

Our model has certain limitations. Although our data do suggest physicians' decision-making is more compatible with dual processing model than with the EUT or a simple regret model (Figure 2), our sample size was not large enough to provide more conclusive support in favor of dual processing model in each specific scenario (Table 3). This was the main limitation of our study. Nevertheless, theoretically, the results fit dual processing theories well, because treatment of PE is familiar to most physicians and AML is not. Novel problems trigger type 2 processing; so, for the relatively unfamiliar AML scenarios, dual processing (which takes both type 1 and type 2 processes into account) has predictive advantage. We should, of course, note that our results do not exclude the possibility that some people do act according to either EUT or regret model (Figure 2). In addition, as noted earlier, there are many dual-processing theories [38] and we evaluated a specific dual-processing model that is applicable to single-point clinical decisions such as those described in the vignettes [6] (see Additional file 1). A different model and experimental design would be needed for testing the way physicians make repeated decisions.

Our results also hold promise in medical education. We demonstrated that, at least in some circumstances, physicians do act according to the threshold model. Therefore, all medical curricula should include the teaching the threshold model(s). Although, on average, dual processing model has performed better, we believe that all 3 models should be taught because they collectively take into account the most salient features of human decision-making (assessment of the likelihood of disease and benefit/harms ratio), which are determined by both type 1 (fast, intuitive) and type 2 (slow, deliberative) reasoning processes. In addition, as outlined above, these descriptive models may conceivably be used in prescriptive fashion under some circumstances. For example, in circumstances where our affect plays a key role in the way we feel the consequences of benefits and harms, we may rely on regret approach. Conversely, where empirical evidence on benefits and harms is a driver of decision-making, then application of EUT may still be more suitable. However, we suspect that integration of both approaches, regret- and EUT-based, into dual processing model will be useful to most users. The details of how this integration may work is beyond a scope of this paper, but is sketched in [6].

Certainly, we need confirmatory and larger studies to reproduce (or refute) our results. While we found that the vignettes were judged by the vast majority of physicians as realistic examples of real-life clinical cases, it is still possible that different scenarios and different wording may elicit different responses. Although including realistic and familiar scenarios can be deemed as one of the strengths of our analysis, it has generated some analytical problems, as outlined above. Therefore, the future research should include larger studies with relatively less familiar, but still realistic-case vignettes.

Conclusions

We find that physicians appear to make treatment decisions according to the threshold model. Furthermore, physicians' decision-making seems more compatible with the dual processing model than with either EUT or a simple regret model. While larger confirmatory studies are needed to affirm our results, the findings of this study may help improve our understanding of clinical decision making under diagnostic uncertainty and may be helpful in development of medical education curricula and practice guidelines.

Additional files

Additional file 1: The survey.

Additional file 2: Table S1. Sensitivity analysis.

Abbreviations

EUT: Expected utility theory; T_{EUT}: Expected utility theory based threshold; T_{RG}: Regret-based threshold; T_{DP}: Dual-processing theory based threshold; B/H: Benefit to harm ratio; PE: Pulmonary embolism; AML: Acute myeloid leukemia; B_{aut}: Automatic benefits assessment; H_{aut}: Automatic harms assessment; B_1: Initial type 1 benefits assessment; H_1: Initial type 1 harms assessment; DVAS: Dual Visual Analog Scale; B_2: Objective benefits assessment; H_2: Objective harms assessment; OR: Odds ratio; CI: Confidence interval.

Competing interests

None of the authors have any financial competing interests to disclose.

Authors' contributions

BD was responsible for concept and design of the study, analysis and interpretation of data, and drafting the manuscript. SE contributed to study design, analysis and interpretation of data, and revision of the manuscript for critically important intellectual content. TR contributed to study design, acquisition of data, analysis and interpretation of data, and revision of the manuscript for critically important intellectual content. IH contributed to analysis and interpretation of data and revision of the manuscript for critically important intellectual content. BM contributed to analysis and interpretation of data and revision of the manuscript for critically important intellectual content. AT contributed to study design, data acquisition, and revision of the manuscript for critically important intellectual content. AK contributed to study design, interpretation of data, and drafting of the manuscript. JB contributed to concept and study design and revision of the manuscript for critically important intellectual content. ST contributed to acquisition of data, and revision of the manuscript for critically important intellectual content. JCB contributed to study design, analysis and interpretation of data, and revision of the manuscript for critically important intellectual content. All authors read and approved the final manuscript.

Acknowledgments

This study was supported in part by the DoD grant #W81 XWH 09-2-0175 (PI: Djulbegovic). We thank Drs. Stephen Pauker and Jef Van den Ende of the Instituut voor tropische geneeskunde, Antwerpen, Belgium for most helpful comments on the earlier versions of this paper. We also are most grateful to Dr. Elizabeth Pathak for help to improve readability of the manuscript from a general readership point of view.

Author details
[1]Department of Internal Medicine, Division of Evidence-based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL, USA. [2]Department of Health Outcomes and Behavior, Moffitt Cancer Center & Research Institute, Tampa, FL, USA. [3]Department of Hematology, Moffitt Cancer Center & Research Institute, Tampa, FL, USA. [4]De Montfort University, Leicester, UK. [5]Indiana University Northwest, Department of Mathematics, Gary, IN, USA. [6]College of Nursing, University of South Florida, Tampa, FL, USA. [7]Center for Advanced Medical Learning & Simulations, University of South Florida, Tampa, FL, USA. [8]USF Health, 3515 East Fletcher Avenue, MDT 1202, Tampa, FL 33612, USA.

Received: 9 July 2013 Accepted: 2 June 2014
Published: 5 June 2014

References

1. Pauker SG, Kassirer J: **The threshold approach to clinical decision making.** *N Engl J Med* 1980, 302:1109–1117.
2. Pauker SG, Kassirer JP: **Therapeutic decision making: a cost benefit analysis.** *N Engl J Med* 1975, 293:229–234.
3. Djulbegovic B, Hozo I, Schwartz A, McMasters K: **Acceptable regret in medical decision making.** *Med Hypotheses* 1999, 53:253–259.
4. Hozo I, Djulbegovic B: **When is diagnostic testing inappropriate or irrational? Acceptable regret approach.** *Med Decis Making* 2008, 28(4):540–553.
5. Hozo I, Djulbegovic B: **Will insistence on practicing medicine according to expected utility theory lead to an increase in diagnostic testing?** *Med Decis Making* 2009, 29:320–322.
6. Djulbegovic B, Hozo I, Beckstead J, Tsalatsanis A, Pauker SG: **Dual processing model of medical decision-making.** *BMC Med Inform Decis Mak* 2012, 12(1):94.
7. Kahneman D: **Maps of bounded rationality: psychology for behavioral economics.** *American Economic Review* 2003, 93:1449–1475.
8. Kahnemen D: *Thinking fast and slow.* New York: Farrar, Straus and Giroux; 2011.
9. Evans JSTBT: *Hypothetical thinking. Dual processes in reasoning and judgement.* New York: Psychology Press: Taylor and Francis Group; 2007.
10. Stanovich KE, West RF: **Individual differences in reasoning: implications for the rationality debate?** *Behav Brain Sci* 2000, 23:645–726.
11. Stanovich KE: *Rationality and the Reflective Mind.* Oxford: Oxford University Press; 2011.
12. Croskerry P: **Clinical cognition and diagnostic error: applications of a dual process model of reasoning.** *Adv Health Sci Educ Theory Pract* 2009, 14(Suppl 1):27–35.
13. Croskerry P: **A universal model of diagnostic reasoning.** *Acad Med* 2009, 84(8):1022–1028.
14. Croskerry P, Abbass A, Wu AW: **Emotional influences in patient safety.** *J Patient Saf* 2010, 6(4):199–205.
15. Croskerry P, Nimmo GR: **Better clinical decision making and reducing diagnostic error.** *J R Coll Physicians Edinb* 2011, 41(2):155–162.
16. Slovic P, Finucane ML, Peters E, MacGregor DG: **Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality.** *Risk Anal* 2004, 24(2):311–322.
17. Tversky A, Kahneman D: **The framing of decisions and the psychology of choice.** *Science* 1981, 211(4481):453–458.
18. Basinga P, Moreira J, Bisoffi Z, Bisig B, Van den Ende J: **Why are clinicians reluctant to treat smear-negative tuberculosis? An inquiry about treatment thresholds in Rwanda.** *Med Decis Making* 2007, 27(1):53–60.
19. Eisenberg JM, Hershey JC: **Derived thresholds: determining the diagnostic probabilities at which clinicians initiate testing and treatment.** *Med Decis Making* 1983, 3:155–168.
20. Moreira J, Alarcon F, Bisoffi Z, Rivera J, Salinas R, Menten J, Duenas G, Van den Ende J: **Tuberculous meningitis: does lowering the treatment threshold result in many more treated patients?** *Trop Med Int Health* 2008, 13(1):68–75.
21. Tuyisenge L, Ndimubanzi CP, Ndayisaba G, Muganga N, Menten J, Boelaert M, Van den Ende J: **Evaluation of latent class analysis and decision thresholds to guide the diagnosis of pediatric tuberculosis in a Rwandan reference hospital.** *Pediatr Infect Dis J* 2010, 29:e11–e18.
22. Zeelenberg M, Pieters R: **A theory of regret regulation 1.1.** *J Consumer Psychol* 2007, 17:29–35.
23. Zeelenberg M, Pieters R: **A Theory of Regret Regulation 1.0.** *J Consumer Psychol* 2007, 17(1):3–18.
24. Jannis IL, Mann L: *Decision Making. A psychological Analysis of Conflict, Choice, and Commitment.* London: The Free Press; 1977.
25. Tsalatsanis A, Hozo I, Vickers A, Djulbegovic B: **A regret theory approach to decision curve analysis: A novel method for eliciting decision makers' preferences and decision-making.** *BMC Med Inform Decis Mak* 2010, 10(1):51.
26. Evans JSTBT: **The heuristic-analytic theory of reasoning: extension and evaluation.** *Psychon Bull Rev* 2006, 13:378–395.
27. Evans JSTBT: *Thinking Twice. Two Minds in One Brain.* Oxford: Oxford University Press; 2010.
28. Mukherjee K: **A dual system model of preferences under risk.** *Psychol Rev* 2010, 177(1):243–255.
29. Evans JSTBT: **Dual-process theories of reasoning: Contemporary issues and developmental applications.** *Developmental Review* 2011, 31:86–102.
30. McNemar Q: **Note on the sampling error of the difference between correlated proportions or percentages.** *Psychometrika* 1947, 12(2):153–157.
31. STATA Corporation: *STATA, ver. 12.* College Station, TX: 2010.
32. Berwick DM, Hackbarth AD: **Eliminating Waste in US Health Care.** *JAMA* 2012, 307(14):1513–1516.
33. Cassel CK, Guest JA: **Choosing Wisely.** *JAMA* 2012, 307(17):1801–1802.
34. Van den Ende J, Moreira J, Tuyisenge L, Bisoffi Z: **An Inquiry About Clinicians' View of the Distribution of Posttest Probabilities: Possible Consequences for Applying the Threshold Concept.** *Med Decis Making* 2013, 33(2):136–8.
35. Djulbegovic B, Paul A: **From efficacy to effectiveness in the face of uncertainty: indication creep and prevention creep.** *JAMA* 2011, 305 (19):2005–2006.
36. Kahneman D, Tversky A: **"Prospect theory": an analysis of decion under risk.** *Econometrica* 1979, 47:263–291.
37. Kahneman D, Wakker PP, Sarin RK: **Back to Bentham? Explorations of Experienced Utility.** *Quarterly Journal of Economics* 1997, 112:375–405.
38. Reyna VF: **A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory.** *Judgment and Decision Making* 2012, 7(3):332–359.
39. Elqayam S: **Grounded rationality: descriptivism in epistemic context.** *Synthese* 2012, 189:39–49.
40. US Preventive Service Task Force: **Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement.** *Ann Intern Med* 2009, 151:716–726.
41. Editors: **When Evidence Collides With Anecdote, Politics, and Emotion: Breast Cancer Screening.** *Ann Intern Med* 2010, 152(8):531–532.

doi:10.1186/1472-6947-14-47
Cite this article as: Djulbegovic *et al.*: How do physicians decide to treat: an empirical evaluation of the threshold model. *BMC Medical Informatics and Decision Making* 2014 14:47.

Defining Optimum Treatment of Patients With Pancreatic Adenocarcinoma Using Regret-Based Decision Curve Analysis

Jonathan M. Hernandez, MD, Athanasios Tsalatsanis, PhD,†‡ Leigh Ann Humphries, BA,* Branko Miladinovic, PhD,†‡ Benjamin Djulbegovic, MD, PhD,†‡§ and Vic Velanovich, MD**

Objective: To use regret decision theory methodology to assess three treatment strategies in pancreatic adenocarcinoma.

Background: Pancreatic adenocarcinoma is uniformly fatal without operative intervention. Resection can prolong survival in some patients; however, it is associated with significant morbidity and mortality. Regret theory serves as a novel framework linking both rationality and intuition to determine the optimal course for physicians facing difficult decisions related to treatment.

Methods: We used the Cox proportional hazards model to predict survival of patients with pancreatic adenocarcinoma and generated a decision model using regret-based decision curve analysis, which integrates both the patient's prognosis and the physician's preferences expressed in terms of regret associated with a certain action. A physician's treatment preferences are indicated by a threshold probability, which is the probability of death/survival at which the physician is uncertain whether or not to perform surgery. The analysis modeled 3 possible choices: perform surgery on all patients; never perform surgery; and act according to the prediction model.

Results: The records of 156 consecutive patients with pancreatic adenocarcinoma were retrospectively evaluated by a single surgeon at a tertiary referral center. Significant independent predictors of overall survival included preoperative stage [$P = 0.005$; 95% confidence interval (CI), 1.19–2.27], vitality ($P < 0.001$; 95% CI, 0.96–0.98), daily physical function ($P < 0.001$; 95% CI, 0.97–0.99), and pathological stage ($P < 0.001$; 95% CI, 3.06–16.05). Compared with the "always aggressive" or "always passive" surgical treatment strategies, the survival model was associated with the least amount of regret for a wide range of threshold probabilities.

Conclusions: Regret-based decision curve analysis provides a novel perspective for making treatment-related decisions by incorporating the decision maker's preferences expressed as his or her estimates of benefits and harms associated with the treatment considered.

Keywords: outcomes, pancreatic cancer, pancreatic cancer survival, pancreatic resection, regret decision analysis

(*Ann Surg* 2014;259:1208–1214)

Although significant progress has be made over the last 2 decades in reducing perioperative mortality for patients with localized pancreatic adenocarcinoma, pancreaticoduodenectomy remains associated with significant morbidity.[1,2] Moreover, long-term survival has remained unchanged and persistently elusive for the vast majority of patients with the disease.[3,4] Operative extirpation, for which about 15% to 20% of patients are eligible, is undertaken when technically feasible because it offers the only opportunity for prolonged survival and because there are few alternative treatments—each of which has limited efficacy.[5] However, even among patients undergoing complete tumor extirpation with negative margins, the disease recurs in 40% of the patients within 6 months, most commonly in the form of liver metastasis.[6] These patients may derive little-to-no survival benefit from local control, while potentially suffering from operative morbidity.[6] Selection of patients likely to benefit from aggressive local control is therefore particularly important in the management of patients with radiographically localized pancreatic adenocarcinoma.

Decision analysis typically defines the probability of an event and provides the optimal model among alternative clinical management strategies, thus maximizing a definable outcome.[7,8] Probability models based on diagnostic and prognostic variables have been used to assist physician decision making regarding various treatments and interventions, including resection for cancer, although the effectiveness of the models remains questionable.[9–15] The reasons behind this skepticism include the probabilistic nature of these models that adds complexity to the decision process and, importantly, the reliance of most of these models on expected utility theory, which is often violated during decision making.[16–20]

We recently developed a decision methodology that overcomes the limitations of probabilistic survival models and that can be used to facilitate medical decisions based on the decision maker's preferences.[19,20] Our methodology, regret-based decision curve analysis or *Regret DCA*, relies on the cognitive emotion of regret to identify conditions under which a physician is unsure about the choice between alternative treatment strategies.[19,20] Surgeons, as with any decision maker, may experience regret (defined as the difference between the utility of an action taken and utility of an alternative action) if they eventually realize that a decision they made was suboptimal and that an alternative form of treatment would have been preferable.[21–27] Regret DCA uses this regret to compute the threshold probability at which the physician is uncertain about which treatment strategy to recommend to his or her patient. In this study, we used Regret DCA to facilitate treatment decisions for a cohort of patients with localized, resectable pancreatic adenocarcinoma.

The intention of this article is to present a novel decision methodology that relies on regret theory and attempts to explain medical decision making for surgeons treating patients with pancreatic adenocarcinoma. Despite the fact that the prediction model presented has been well fitted to our data, its role in this article is secondary and its purpose is to demonstrate how the regret methodology can be used to evaluate 3 management strategies: aggressive, passive, or model-based decision making. In this context, we have demonstrated that the prediction model performs better the other 2 strategies in terms of regret.

MATERIALS AND METHODS

The records of 156 consecutive patients referred for surgical consultation from January 2005 to 2009 with pancreatic adenocarcinoma were retrospectively reviewed by a single surgeon at a tertiary referral center. The diagnosis was confirmed by histological evaluation, and disease stage was determined both by pathological

From the *Department of Surgery, Division of General Surgery, University of South Florida, Tampa, FL; †Center for Evidence-Based Medicine, University of South Florida, Tampa, FL; ‡Department of Internal Medicine, Division of Evidence-Based Medicine, Tampa, FL; and §Department of Hematology and Health Outcomes and Behavior, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL.

Disclosure: Partially supported by the Department of Army grant W81 XWH 09-2-0175. The authors declare no conflicts of interest.

Reprints: Vic Velanovich, MD, Department of General Surgery, 1 Tampa General Circle, F145, Tampa, FL 33606. E-mail: vvelanov@health.usf.edu.

ISSN: 0003-4932/13/25906-1208
DOI: 10.1097/SLA.0000000000000310

evaluation of the resected specimen and by imaging. All patients had been administered the SF-36 Health Survey to assess quality of life, which includes 36 statements grouped into 8 domains of quality of life: physical functioning, physical role, bodily pain, general health, vitality, social functioning, emotional role, and mental health. The SF-36 uses a Likert scale of 0 to 100, with higher scores indicating better/normal health or physical functioning. It has been previously demonstrated that the SF-36 correlates well with pathology, survival, stage, and resectability of pancreatic lesions.[28]

The distribution for overall survival was estimated using the Kaplan-Meier method. Cox proportional hazards modeling was used to determine the effect on survival of the following 12 covariates, including those described by SF-36: age, sex, stage, adjuvant therapy, physical functioning, role-physical, role-emotional, bodily pain, pretreatment vitality, mental health, social functioning, and general health. Additional covariates such as tumor characteristics (lymphovascular invasion, perineural invasion, etc) could potentially influence the output of the Cox model; however, this information is typically unknown to the surgeon a priori. Furthermore, such covariates were not included in the analysis because our data set was originally constructed on the basis of the methods and protocols designed for a study[28] focusing on the quality of life, pathology, resectability, and survival in patients with pancreatic lesions. The model was created using stepwise elimination on all variables ($P < 0.15$ to enter, and $P < 0.20$ to stay). The proportional hazards assumption was examined using Schoenfeld residuals. The importance of each variable and the discriminative ability of the Cox model were examined using the Royston-Sauerbrei discrimination statistic D and explained variation R^2_D.[29] All continuous variables were centered about the mean. All analyses were performed using STATA.[30]

To derive the optimal treatment strategy, we then used the Regret DCA methodology.[19,20] Regret DCA uses the decision maker's feeling of regret to compute the threshold probability at which he or she is uncertain about alternative actions, for example, to operate or not to operate. In considering decisions for patients with pancreatic adenocarcinoma, we considered survival less than 7 months from the time of tumor extirpation as being unlikely to have imparted a survival advantage and therefore unnecessary based upon median survival of patients with locally advanced, nonmetastatic disease.[31] On the basis of this assumption, we formulated a decision model that compares an individual patient's prognosis with the threshold probability at which the surgeon would be indifferent about recommending surgery.

Typically, decision theory suggests that a person should be treated if the probability of an event (ie, the patient develops a disease; the patient dies; the patient survives longer than a predefined time frame, etc) is greater than or equal to a threshold probability.[7,8,32] In this article, we sought to treat the patients who were likely to survive longer than 7 months from the time of their resection. Therefore, the convention used is as follows: if the patient's probability of surviving 7 months is greater than or equal to the threshold probability ($s \geq P_t$), the surgeon should offer resection. If the patient's probability of survival is less than the threshold probability ($s < P_t$), the patient may be unlikely to benefit substantially from surgery and the surgeon should not recommend resection in favor of medical alternatives.

The probability of survival can be computed for each patient on the basis of the Cox survival model previously described. However, the threshold probability is subject to each surgeon's preferences and clinical practice attitudes. At the individual level, it can be computed as follows[19,20]:

$$P_t = \frac{1}{1 + \frac{\text{Regret of omission}}{\text{Regret of comission}}} \quad (1)$$

We define "regret of omission" as the regret felt by a surgeon who withheld necessary surgery from a patient who may have benefited from that resection (patients with localized disease who lived longer than 7 months). Conversely, "regret of commission" is the regret felt by a surgeon who performed an unnecessary surgery on a patient who derived no benefit from that operation (eg, the patient died as a result of the procedure or died within 7 months from the time of resection). Both regret values can be determined using the dual visual analogue scales (DVAS) (Fig. 1).[19,20] Formally, regret can be expressed as the difference between the utility of the outcome of an action taken and the utility of the outcome of the action that, in retrospect, should have been taken.[21–27] Commonly used techniques for estimating utility, and therefore decision maker's preferences, such as standard gamble and time tradeoff are time-consuming, cognitively complex, and have been shown to lead to biased estimates of people's preferences.[33–35] Instead, in this article, we use the DVAS to estimate directly the values of regret of commission and omission.[19,20] The DVAS comprise two 100-point scales, each anchored to no regret and maximum regret. One of the scales is used to elicit regret of omission and the other to elicit regret of commission (Fig. 1).

After computing the surgeon's threshold probability, the clinical question regarding treatment of patients with pancreatic

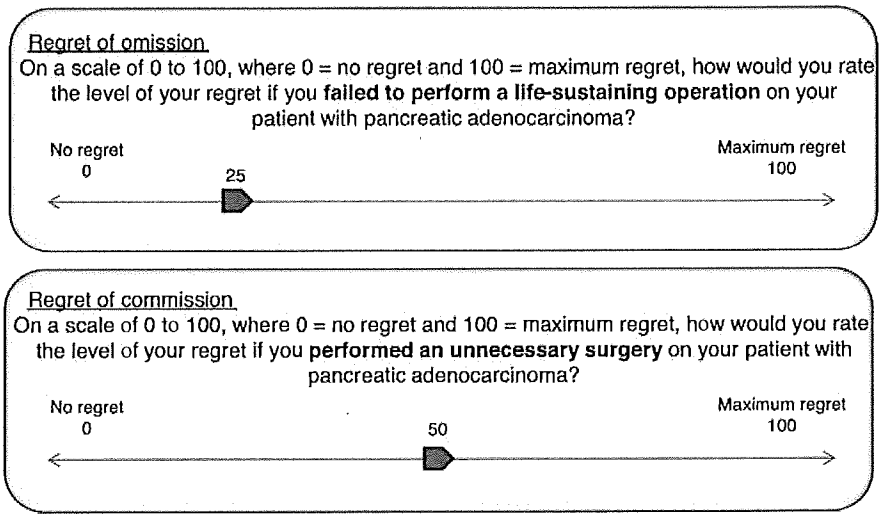


FIGURE 1. Dual visual analog scales. These scales are used for the elicitation of the decision maker's threshold probability. The questions depicted are case specific.

adenocarcinoma can be broken down into 3 strategies: (1) surgeons can stay passive and allow the disease to run its course; (2) surgeons can be aggressive and recommend resection on all patients; or (3) surgeons can use prediction model for guidance. Any of these strategies may cause regret if the outcome is poor. Under the Regret DCA methodology, the optimal strategy is the one that will cause the least amount of regret if that strategy is proven suboptimal. Formally, regret can be expressed as the difference between the utility of the outcome of the action taken and the utility of the outcome of the action that, in retrospect, should have been taken.[21–27] Considering the decision tree that describes this clinical problem (Fig. 2), we can compute the expected regret associated with each of the 3 strategies as follows:

$$\mathrm{ERg[No\ surgery]} = (1 - s) \times \frac{P_t}{1 - P_t} \tag{2}$$

$$\mathrm{ERg[Surgery]} = s \tag{3}$$

$$\mathrm{ERg[Model]} = \frac{\#\mathrm{FP}}{n} \times \frac{P_t}{1 - P_t} + \frac{\#\mathrm{FN}}{n} \tag{4}$$

The values of #FP and #FN correspond to the number of false-positive (FP) and false-negative (FN) results, respectively, as compared with the actual patient outcomes used for the development of the prediction model, and the number of patients in the data set is n. We define true-positive (TP), true-negative (TN), FP, and FN results as follows:

TP: The number of patients who will survive longer than 7 months and for whom the estimated probability of survival is greater than or equal to the threshold probability (ie, the patients who should receive surgery).
TN: The number of patients who will die in 7 months and for whom the estimated probability of survival is less than the threshold probability (ie, the patients who should *not* receive surgery).
FP: The number of patients who will die within 7 months and for whom the estimated probability of survival is greater than or equal to the threshold probability (ie, the patients who received unnecessary surgery).
FN: The number of patients who will survive longer than 7 months and for whom the estimated probability of survival is less than the threshold probability (ie, the number of patients who should have received surgery but did not).

As shown in equations 2 and 4, the expected regret associated with each strategy is a function of the physician's threshold probability. To identify the least regretful action, the Regret DCA methodology computes the expected regret for a range of threshold probabilities (0–100), and expected regret is then graphed against the threshold probability for each of the 3 actions. The action with the lowest value of expected regret corresponds to the most desired action, given a certain threshold probability.

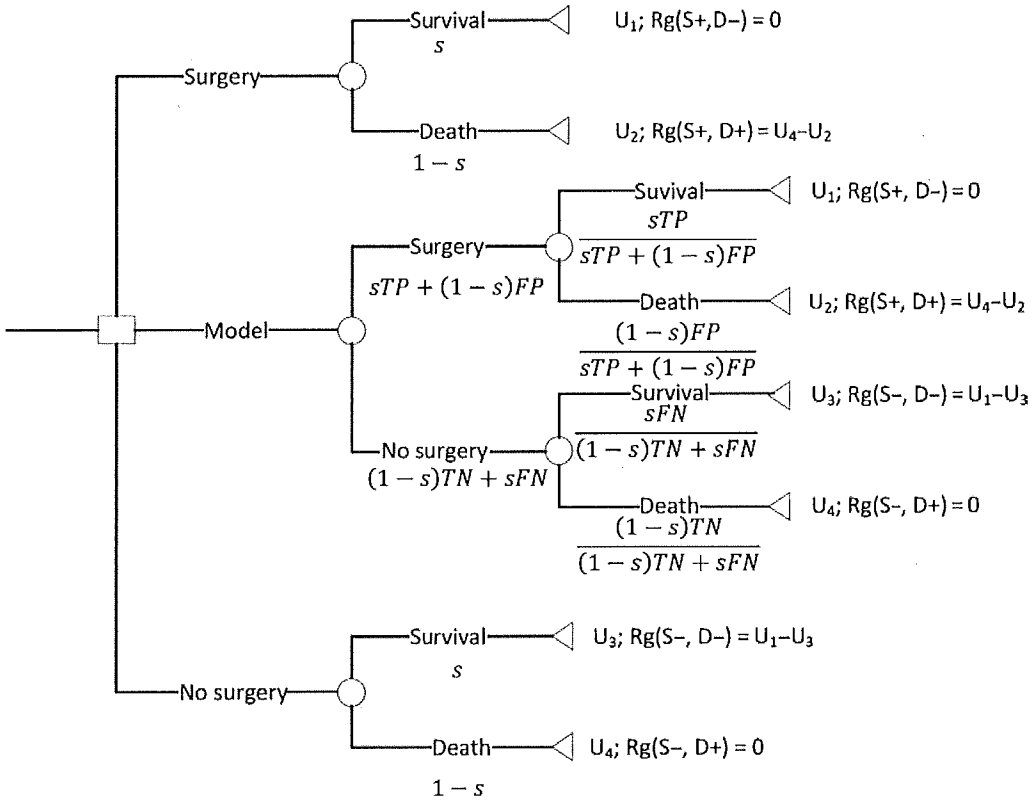


FIGURE 2. Decision model for performing surgery on patients suffering from pancreatic adenocarcinoma. s denotes the probability of survival, $S \pm$ denotes surgery or no surgery, $D \pm$ denotes death or no death, U_i are the utilities associated with each outcome, and Rg is the regret associated with each action. For example, Rg($S-$, $D+$) is the regret associated with not performing a surgery for a patient who died within 7 months.

RESULTS

Patient Characteristics

A total of 156 patients with histologically confirmed primary pancreatic adenocarcinoma were included. The mean age was 65.9 ± 10 years, 83% were stage I or II, 54% were resected, 66% received chemotherapy, and the median survival was 18 months [95% confidence interval (CI), 12–26] (mean survival was 15.7 ± 25 months). The SF-36 scores revealed that role-physical and pretreatment vitality had the lowest scores and mental health had the highest score (Table 1). The distribution of overall survival is shown in Figure 3.

Survival Model

Of the 12 variables included in the data set, 3 met the stepwise inclusion criteria and were used to construct the survival model: stage, pretreatment vitality, and role-physical (daily physical functioning). The explained variation of the fitted model was $R^2_D = 0.4$ (95% CI, 0.27–0.52), and the proportional hazard assumption were not violated ($P < 0.96$). Table 2 presents the estimates of hazard ratio for the Cox prediction model.

TABLE 1. Patient Demographics and SF-36 Scores

Male:female, n (%)	70 (45):86 (55)
Age, yr	65.9 ± 10
Stage: n (%)	
I	61 (39)
II	68 (44)
III	25 (16)
0	2 (1)
SF-36 scores*	
Physical functioning	55.2 ± 31
Role-physical	35.5 ± 44
Role-emotional	57.4 ± 46
Bodily pain	55.5 ± 30
Pretreatment vitality	41.8 ± 24
Mental health	70.3 ± 21
Social functioning	60.8 ± 31
General health	60.7 ± 22
Patients undergoing resection, n (%)	85 (54)
Patients receiving chemotherapy, n (%)	103 (66)
Survival, mo	15.7 ± 25

Values are the mean ± SEM unless otherwise indicated.
*SF-36 Health Survey, rated from 0 to 100 on a Likert scale, with higher scores indicating better health or physical function (ref).

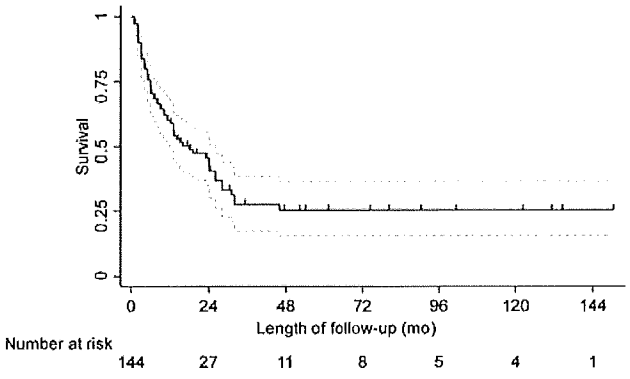


FIGURE 3. Overall survival of patients with pancreatic adenocarcinoma expressed as Kaplan-Meier survival and 95% CI bands. Vertical bars (|) denote censored observations.

Regret Decision Curve Analysis

We used Regret DCA to evaluate the 3 management strategies: (1) recommend against potentially curative surgery in favor chemotherapy or chemoradiotherapy; (2) be aggressive and recommend resection; and (3) use the prediction model as a decision aid. Figure 4 depicts the expected regret as a function of threshold probability for each of the 3 management strategies. As shown, the least regretful strategy for threshold probabilities greater than 5% is to use the prediction model. For threshold probabilities between 80% and 87%, the regret curve associated with the prediction model is subject to noise[36] that we attribute to the error term of the Cox prediction model. We assume that the prediction model remains the least regretful strategy within the 80% to 87% range as well. Our results demonstrate that the survival model has significant clinical value for the majority of decision makers.

Hypothetical Case Study

A 72-year-old woman with diabetes and hypertension has a diagnosis of pancreatic adenocarcinoma after undergoing endoscopic

TABLE 2. Hazard Ratio Estimates of the Prediction Model

	Hazard Ratio	$P > \|z\|$	95% CI
Stage	1.99	0.001	1.32–2.99
Pretreatment vitality	0.98	0.030	0.97–0.99
Role-physical	0.98	0.005	0.98–0.99

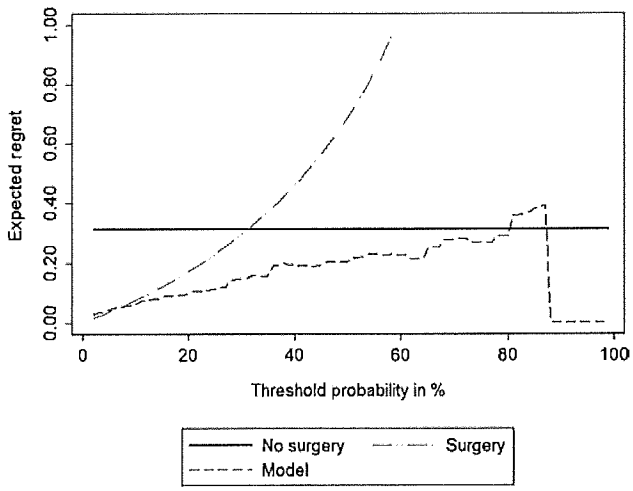


FIGURE 4. Regret DCA for the survival model constructed using Cox regression on 3 variables. Dashed and dotted line denotes the decision to perform surgery; solid line denotes the decision not to perform surgery on any patient; and dashed line denotes the use of the survival model to perform surgery. The optimal strategy is the action that results in the least amount of regret in case it is proven wrong. For threshold probabilities of 0% to 5%, the optimal strategy is to perform surgery on all patients, whereas for threshold probabilities greater than 5% the optimal strategy is to consult the survival model. For threshold probabilities between 80% and 87%, the regret curve associated with the prediction model is subject to noise associated with the error of the prediction model; therefore, we assume that the prediction model remains the least regretful strategy.

retrograde cholangiopancreatography and common bile duct stenting for obstructive jaundice. She is currently without pain and is tolerating a regular diet. Her jaundice resolved after the placement of her biliary stent. Her computed tomographic scan demonstrates a localized mass in the head of the pancreas without involvement of the superior mesenteric vein, portal vein, superior mesenteric artery, or hepatic arteries. The patient is active and able to perform all activities of daily living. She expresses a strong desire to spend as much time as she can with her grandchildren.

We demonstrate the decision process assuming 2 types of hypothetical decision makers: one surgeon is extremely selective in offering resection to patients with pancreatic adenocarcinoma (surgeon 1), and the second surgeon (surgeon 2) generally offers resection to all patients with radiographically resectable disease. The process, depicted in Figure 5, is initiated with the elicitation of the surgeon's preferences. Using the DVAS method (Fig. 1), we estimate the threshold probability as a function of regret of omission and regret of commission (equation 1). Suppose that the answers to the questions shown in Figure 1 for the surgeons are as follows:

Surgeon 1: Regret of omission: 20; regret of commission: 90. Therefore, the threshold probability is equal to 81.8% (equation 1).
Surgeon 2: Regret of omission: 90; regret of commission: 4. Therefore, the threshold probability is equal to 4.2%.

On the basis of results of Regret DCA (Fig. 4), the optimal and least regretful strategy for surgeon 1 is to use the prognostication model we developed, described earlier. If the patient's estimated probability of survival is greater than or equal to 81.8% (the threshold for surgeon 1), then the optimal strategy is to treat (perform the operation). If the probability of survival is less than 81.8%, then the optimal strategy is to offer alternative treatments (forego resection). Conversely, for surgeon 2, whose threshold probability is equal to 4.2%, the optimal and least regretful strategy is to offer resection.

As mentioned earlier, the Regret DCA methodology can also be used by the patients.[19] For completeness, we present how this process could work. The patient would be asked questions similar to those depicted in Figure 1. We have previously shown that patient ratings of utility scores closely correlate with quality of life after pancreaticoduodenectomy; moreover, this patient-centered assessment many change over time as quality of life improves.[37]

Regret of omission: On a scale of 0 to 100, where 0 = no regret and 100 = maximum regret you could feel, how would you rate your level of regret if you did not have an operation that could have extended your life?
Regret of commission: On a scale of 0 to 100, where 0 = no regret and 100 = maximum regret you could feel, how would you rate your level of regret if you had an operation that did not extend your life?

DISCUSSION

We describe the theory and application of regret decision curve analysis as it applies to surgeons and to decisions regarding operative intervention in patients with pancreatic adenocarcinoma. To the best of our knowledge, this is the first application of Regret DCA to assist surgeons in decision making for patients with pancreatic malignancies. Our approach promotes personalized patient care by incorporating decision maker's preferences from the perspective of regret by estimating a threshold probability for a decision maker. We believe the decision regarding resection for patients with pancreatic adenocarcinoma is particularly well suited for a regret-based approach, given the generally fatal prognosis for this disease, regardless of the decision made.

Modern cognitive theories seek to balance risks and benefits in the decision-making process by taking into account both intuition and analytical processes.[38] We believe that rational decision making should take into account both the formal principles of rationality and human intuition. We have accomplished this using regret, a cognitive emotion, to serve as the link between intuition and analytical thinking.[19,20] Eliciting surgeons' preferences by using regret is likely to prove superior to using traditional utility theory because regret explicitly forces the surgeon to consider consequences of decisions. Our method relies on elicitation of a threshold probability, which must be calculated for every decision maker. In other words, our model forces surgeons to consider the possible outcomes of recommending pancreaticoduodenectomy rather than simply recommending resection for all tumors that appear resectable on radiographic imaging.

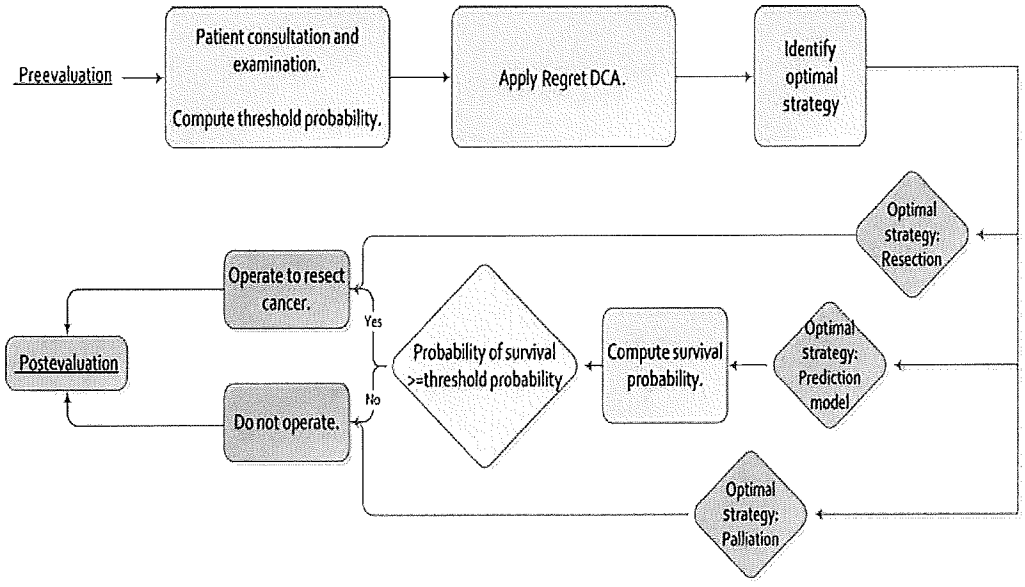


FIGURE 5. Schematic representation of the decision model.

We argue that our approach contributes to the field of decision making, but we acknowledge that it is not a panacea. We do, however, believe that our methodology is best suited for medical decision making primarily associated with tradeoffs between quality and quantity of life. Pancreatic adenocarcinoma meets this criterion: surgical resection may offer an additional year of survival, albeit with the potential for serious morbidity, particularly if the resection is undertaken at low-volume centers.[39,40] For the fortunate 15% to 20% of patients with radiographically localized disease amenable to resection, the median survival ranges from 17 to 23 months.[41] At high-volume institutions with extensive experience, the mortality rate is less than 3% to 5%, but morbidity remains problematic, with early postoperative complication rates of approximately 30% to 40%.[6] Perioperative morbidity and mortality rates recorded in national databases, which include data from a broad spectrum of hospitals and surgeons' experiences, report significantly higher numbers of complications than high-volume tertiary referral centers.[39] The superior results obtained at high-volume tertiary centers may be related to the amount of clinical support that a particular surgeon has at a particular hospital.[42] Applying our model of regret theory may indirectly motivate each surgeon to consider his or her own results with the procedure and to consider the support available within the institution where the procedure is planned when contemplating the best course of action for each patient, further personalizing care.

A significant proportion of patients undergoing resection develop early metastatic disease and have very limited survival and thus derive no benefit from the operative intervention (ie, there is no tradeoff improvement in quality of life). This issue has been addressed with the use of refined definitions of borderline resectability and the use of neoadjuvant therapy.[43] Specifically, this minimally effective chemotherapy, which offers virtually no hope of eradicating disease and little if any therapeutic efficacy, does provide a "window of observation," during which distant metastatic disease may appear and thus spare the patient unnecessary surgery. This approach may minimize regret and results in better overall survival for patients who ultimately are undergoing resection,[44] but it has not been widely adopted across the country or even across academic centers. Similarly, regret theory remains severely underutilized in the health care arena, despite considerable conceptual and empiric interest in its applicability, and in the strong influence of regret on physician decision making.[32,45–47] The lack of incorporation of regret theory into health care delivery is particularly perplexing, especially considering that all medical decisions are accompanied by varying degrees of risk and uncertainty and therefore potential regret. Moreover, recent work has suggested that physicians' behavior can often be explained by regret avoidance,[48] which further substantiates the need to incorporate regret modeling into health care decisions.

As with any novel theoretical work, our application of regret theory to pancreatic adenocarcinoma has limitations. First, we applied the theory retrospectively with assigned cutoff survival values. We assumed maximal regret to be associated with operating on a patient who died within the first 7 months after resection. Excluding death as a result of the procedure (perioperative death), which is always associated with regret, death within 7 months may not necessarily be associated with regret. For example, a patient may have died of an unrelated stroke that could not have been foreseen before resection. Second, our approach has not yet been empirically tested and the prediction model has not been externally validated. Third, the methodology, as presented, is appropriate for point decision making and not necessarily for decisions that reoccur over time—as frequently happens in patient care. Finally, we assumed that there is a single decision maker involved in the process where, in actual practice, a multidisciplinary team of health care providers is involved in treatment decisions.

CONCLUSIONS

We have described a novel approach to surgical decision making using the cognitive emotion of regret, which seeks to personalize care. The goal of our work is to power a computerized decision support tool to assist physicians and patients in making better medical decisions. We envision the tool to be shared by both the physician and the patient during consultation, in which the physician elicits the patient's preferences toward alternative management strategies.

ACKNOWLEDGMENTS

The authors appreciate the assistance of Jane Carver of the University of South Florida Clinical and Translational Science Institute in reviewing the manuscript.

REFERENCES

1. McPhee JT, Hill JS, Whalen GF, et al. Perioperative mortality for pancreatectomy: a national perspective. *Ann Surg.* 2007;246:246–253.
2. Vollmer CM, Sanchez N, Gondek S, et al. A root-cause analysis of mortality following major pancreatectomy. *J Gastrointest Surg.* 2012;16:89–102.
3. Baxter NN, Whitson BA, Tuttle TM. Trends in the treatment and outcome of pancreatic cancer in the United States. *Ann Surg Oncol.* 2007;14:1320–1326.
4. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2008. *CA Cancer J Clin.* 2008;58:71–96.
5. Cress RD, Yin D, Clarke L, et al. Survival among patients with adenocarcinoma of the pancreas: a population-based study (United States). *Cancer Causes Control.* 2006;17:403–409.
6. Beger HG, Rau B, Gansauge F, et al. Treatment of pancreatic cancer: challenge of the facts. *World J Surg.* 2003;27:1075–1084.
7. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. *New Engl J Med.* 1975;293:229–234.
8. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *New Engl J Med.* 1980;302:1109–1117.
9. Abramson MA, Pandharipande P, Ruan D, et al. Radical resection for T1b gallbladder cancer: a decision analysis. *HPB.* 2009;11:656–663.
10. Aloia TA, Fahy BN. A decision analysis model predicts the optimal treatment pathway for patients with colorectal cancer and resectable synchronous liver metastases. *Clin Colorectal Cancer.* 2008;7:197–201.
11. Kattan MW, Cowen ME, Miles BJ. A decision analysis for treatment of clinically localized prostate cancer. *J Gen Intern Med.* 1997;12:299–305.
12. Kulkarni GS, Finelli A, Fleshner NE, et al. Optimal management of high-risk T1G3 bladder cancer: a decision analysis. *PLoS Med.* 2007;4:e284.
13. Lotan Y, Cadeddu JA, Lee JJ, et al. Implications of the prostate cancer prevention trial: a decision analysis model of survival outcomes. *J Clin Oncol.* 2005;23:1911–1920.
14. Steyerberg EW, Marshall PB, Jan Keizer H, et al. Resection of small, residual retroperitoneal masses after chemotherapy for nonseminomatous testicular cancer. *Cancer.* 2000;85:1331–1341.
15. Telford JJ, Saltzman JR, Kuntz KM, et al. Impact of preoperative staging and chemoradiation versus postoperative chemoradiation on outcome in patients with rectal cancer: a decision analysis. *J Natl Cancer Inst.* 2004;96:191–201.
16. Baron J. *Thinking and Deciding.* Cambridge: Cambridge University Press; 2000.
17. Bell DE, Raiffa H, Tversky A. *Decision Making: Descriptive, Normative, and Prescriptive Interactions.* Cambridge: Cambridge University Press; 1988.
18. Dawes RM, Kagan J. *Rational Choice in an Uncertain World.* New York: Harcourt Brace Jovanovich; 1988.
19. Tsalatsanis A, Barnes LE, Hozo I, et al. Extensions to regret-based decision curve analysis: an application to hospice referral for terminal patients. *BMC Med Inform Decis Mak.* 2011;11:77.
20. Tsalatsanis A, Hozo I, Vickers A, et al. A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making. *BMC Med Inform Decis Mak.* 2010;10:51.
21. Djulbegovic B, Hozo I. When should potentially false research findings be considered acceptable? *PLoS Med.* 2007;4:e26.
22. Djulbegovic B, Hozo I, Schwartz A, et al. Acceptable regret in medical decision making. *Med Hypotheses.* 1999;53:253–259.
23. Hozo I, Djulbegovic B. When is diagnostic testing inappropriate or irrational? Acceptable regret approach. *Med Dec Making.* 2008;28:540–553.

24. Hozo I, Djulbegovic B. Will insistence on practicing medicine according to expected utility theory lead to an increase in diagnostic testing? *Med Dec Making*. 2009;29:320–322.
25. Bell DE. Regret in decision making under uncertainty. *Oper Res*. 1982;30: 961–981.
26. Loomes G, Sugden R. Regret theory: an alternative theory of rational choice. *Econ J*. 1982;92:805–824.
27. Zeelenberg M, Pieters R. A theory of regret regulation 1.1. *J Consumer Psychol*. 2007;17:29–35.
28. Velanovich V. The association of quality of life measures with malignancy and survival in patients with pancreatic pathology. *Pancreas*. 2011;40:1063–1069.
29. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med*. 2004;23:723–748.
30. StatCorp. *Stata Statistical Software*. Release 12. College Station, TX: StataCorp LP; 2011.
31. Cohen SJ, Dobelbower R, Lipsitz S, et al. A randomized phase III study of radiotherapy alone or with 5-fluorouracil and mitomycin-C in patients with locally advanced adenocarcinoma of the pancreas: Eastern Cooperative Oncology Group study E8282. *Int J Radiat Oncol Biol Phys*. 2005;62:1345–1350.
32. Djulbegovic B, Hozo I, Lyman GH. Linking evidence-based medicine therapeutic summary measures to clinical decision analysis. *Med Gen Med*. 2000;2:E6.
33. Hunink MM, Glasziou PP, Siegel JE, et al. *Decision Making in Health and Medicine: Integrating Evidence and Values*. Cambridge: Cambridge University Press; 2001.
34. Lichtenstein S, Slovic P. *The Construction of Preference*. New York: Cambridge University Press; 2006.
35. Stiggelbout AM, De Haes J. Patient preference for cancer therapy: an overview of measurement approaches. *J Clin Oncol*. 2001;19:220–230.
36. Vickers AJ, Cronin AM, Elkin EB, et al. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak*. 2008;8:53.
37. Warnick SJ, Velanovich V. Correlation of patient derived utility values and quality of life after pancreaticoduodenectomy for pancreatic cancer. *J Am Coll Surg*. 2006;202:906–911.
38. Kahneman D. Maps of bounded rationality: Psychology for behavioral economics. *Am Econ Rev*. 2003;93:1449–1475.
39. Birkmeyer JD, Finlayson SRG, Tosteson ANA, et al. Effect of hospital volume on in-hospital mortality with pancreaticoduodenectomy. *Surgery*. 1999;125:250–256.
40. Gouma DJ, Van Geenen RCI, van Gulik TM, et al. Rates of complications and death after pancreaticoduodenectomy: risk factors and the impact of hospital volume. *Ann Surg*. 2000;232:786–795.
41. Rudloff U, Maker AV, Brennan MF, et al. Randomized clinical trials in pancreatic adenocarcinoma. *Surg Oncol Clin North Am*. 2010;19:115–150.
42. Joseph B, Morton JM, Hernandez-Boussard T, et al. Relationship between hospital volume, system clinical resources and mortality in pancreatic resection. *J Am Coll Surg*. 2009;208:520–527.
43. Lal A, Christians K, Evans DB. Management of borderline resectable pancreatic cancer. *Surg Oncol Clin N Am*. 2010;19:359–370.
44. Katz MHG, Pisters PWT, Evans DB, et al. Borderline resectables pancreatic cancer: the importance of this emerging stage of disease. *J Am Coll Surg*. 2008;206:833–846.
45. Coricelli G. The potential role of regret in the physician-patient relationship: insights from neuroeconomics. *Adv Health Econ Health Serv Res*. 2008;20: 85–97.
46. Djulbegovic B, Hozo I, Beckstead J, et al. Dual processing model of medical decision-making. *BMC Med Inform Decis Mak*. 2012;12:94.
47. Freemantle N. Are decisions taken by health care professionals rational? A non systematic review of experimental and quasi experimental literature. *Health Policy*. 1996;38:71–81.
48. Courvoisier DS, Agoritsas T, Perneger TV, et al. Regrets associated with providing healthcare: qualitative study of experiences of hospital-based physicians and nurses. *PLoS One*. 2011;6:e23138.

Evaluation of Physicians' Cognitive Styles

Benjamin Djulbegovic, MD, PhD, Jason W. Beckstead, PhD, Shira Elqayam, PhD, Tea Reljic, BS, Iztok Hozo, PhD, Ambuj Kumar, MD, MPH, Janis Cannon-Bowers, PhD, Stephanie Taylor, MD, Athanasios Tsalatsanis, PhD, Brandon Turner, PhD, Charles Paidas, MD

Background. *Patient outcomes critically depend on accuracy of physicians' judgment, yet little is known about individual differences in cognitive styles that underlie physicians' judgments. The objective of this study was to assess physicians' individual differences in cognitive styles relative to age, experience, and degree and type of training.* ***Methods.*** *Physicians at different levels of training and career completed a web-based survey of 6 scales measuring individual differences in cognitive styles (maximizing v. satisficing, analytical v. intuitive reasoning, need for cognition, intolerance toward ambiguity, objectivism, and cognitive reflection). We measured psychometric properties (Cronbach's α) of scales; relationship of age, experience, degree, and type of training; responses to scales; and accuracy on conditional inference task.* ***Results.*** *The study included 165 trainees and 56 attending physicians (median age 31 years; range 25–69 years). All 6 constructs showed acceptable psychometric properties. Surprisingly, we found significant negative correlation between age and satisficing (r = –0.239; P = 0.017). Maximizing (willingness to engage in alternative search strategy) also decreased with age (r = –0.220; P = 0.047). Number of incorrect inferences negatively correlated with satisficing (r = –0.246; P = 0.014). Disposition to suppress intuitive responses was associated with correct responses on 3 of 4 inferential tasks. Trainees showed a tendency to engage in analytical thinking (r = 0.265; P = 0.025), while attendings displayed inclination toward intuitive-experiential thinking (r = 0.427; P = 0.046). However, trainees performed worse on conditional inference task.* ***Conclusion.*** *Physicians capable of suppressing an immediate intuitive response to questions and those scoring higher on rational thinking made fewer inferential mistakes. We found a negative correlation between age and maximizing: Physicians who were more advanced in their careers were less willing to spend time and effort in an exhaustive search for solutions. However, they appeared to have maintained their "mindware" for effective problem solving.* ***Key words:*** *physicians' cognitive styles; individual differences in decision-making; medical decision-making; dual processing theories.* ***(Med Decis Making 2014;34:627–637)***

Received 27 August 2013 from Department of Internal Medicine, Division of Evidence-Based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL (BD, TR, AK, ST, AT); Departments of Hematology and Health Outcomes and Behavior, Moffitt Cancer, Tampa, FL (BD); College of Nursing, University of South Florida, Tampa, FL (JWB); School of Applied Social Sciences, De Montfort University, Leicester, UK (SE); Department of Mathematics, Indiana University Northwest, Gary, IN (IH); Center for Advanced Medical Learning & Simulations, University of South Florida, Tampa, FL (JCB); Department of Psychology, Stanford University, Stanford, CA (BT); and Department of Pediatric Surgery, University of South Florida, Tampa, FL (CP). Financial support for this study was provided in part by a support from University of South Florida Graduate Medical Education. The funding agreement ensured the authors' independence in designing the study, interpreting the data, and writing and publishing the report. The following author is employed by the sponsor: Charles Paidas. The other authors have no conflicts of interest to disclose relevant to the current study. Revision accepted for publication 4 February 2014.

DOI: 10.1177/0272989X14525855

Each day, physicians make numerous decisions that have crucial consequences for their patients. Decisions made by physicians are widely recognized as being affected by 3 sets of characteristics: 1) decision features, that is characteristics of the decision itself; 2) situational factors; and 3) individual differences among decision makers.[1] Historically, researchers have focused on studying the first two sets of factors. No study to date has attempted to comprehensively understand individual differences in physicians' cognitive styles. By *cognitive style* we mean a propensity to favor one decision-making or reasoning approach over

Supplementary material for this article is available on the *Medical Decision Making* Web site at http://mdm.sagepub.com/supplemental.

Address correspondence to Benjamin Djulbegovic, USF Health, 12901 Bruce B. Downs Boulevard, MDC27, Tampa, FL 33612; e-mail: bdjulbeg@health.usf.edu.

another. Understanding the differences in individual cognitive styles is important for both physicians and patients as it can lead to prescriptive interventions that may improve decision making.

Investigating differences in cognitive styles has been popular in the cognitive sciences for decades and has flourished in higher mental processing research in more recent times.[2]

Numerous constructs for explaining individual differences in decision making have been proposed over the years, and in some cases instruments have been developed to quantify important individual differences. Our review of decision-making literature has led us to consider 6 key constructs on which physician decision makers are likely to show important individual differences. These are 1) maximizing and satisficing, which refer to the amount of effort an individual is willing to expend to determine the absolute best option in a decision[3]; 2) need for cognition, or the degree to which individuals prefer to engage in and derive enjoyment from cognitive activities[4]; 3) intolerance for ambiguity, which refers to an individual's ability to feel comfortable and accept situations where variables, alternatives, or outcomes are poorly defined or unclear[5]; 4) objectivism, or the tendency to seek empirical information under conditions of uncertainty and to attempt to process it in a rational and logical fashion[6]; 5) cognitive reflection, that is, the ability or disposition to resist reporting the response that first comes to mind[7]; and 6) propensity toward intuitive-experiential versus analytical-rational thinking,[8] which can be explained by dual process theories of cognition.[9,10]

Despite some challenges,[11] dual processing theories have increasingly been accepted as a dominant account of cognitive processes that characterize human decision making.[9] One of the central effects highlighted by dual processing theories of reasoning is "belief bias"—the tendency to evaluate the validity of an argument on the basis of whether one agrees with the contents rather than on whether the conclusion follows logically from the premises.[12–14] The previous studies in samples from the general population showed that high-ability participants (i.e., experts) have more counter-examples accessible to them and are less influenced by belief bias compared with low-ability participants (i.e., trainees-novices).[15] In a medical context, this observation leads to the hypothesis that faculty and attending physicians ("experts") are expected to perform better on reasoning clinical tasks than are residents and fellows ("trainees-novices"). However, it is not clear that experts and novices rely on the same or different cognitive mechanisms. It is also not clear whether the propensity for using one decision-making style over another (e.g., satisficing v. maximizing, intuitive-experiential v. analytic-deliberative, etc) is affected by the type of training (cognitive-oriented v. procedure-oriented disciplines) or by demographic characteristics such as age and gender. Because such styles are affected by cultural transmission,[16] we would expect them to be sensitive to both expertise and demographic variability.

To address these questions, we assessed the validity of physicians' inferences by measuring cognitive styles and their relationship to accuracy on a conditional inference task. We also sought to understand the extent to which physicians' cognitive styles can be categorized as maximizing versus satisficing or intuitive-experiential versus analytic-deliberative in relation to age, experience, degree (trainee v. attending physicians), and type of training (surgical v. nonsurgical disciplines). Finally, we determined a relationship between these constructs and other constructs of importance to understand individual differences in decision making (such as need for cognition, intolerance for ambiguity, and the disposition to resist reporting the response that first comes to mind).

METHODS

All residents, fellows, and physicians affiliated with the University of South Florida were invited by e-mail to participate in the study. The web-based survey included questions on demographics (i.e., age, level of training, gender, specialty), well-validated scales measuring 6 key constructs on which physician decision makers are likely to show important individual differences, and 4 types of conditional inference (explained below). The complete survey is provided in the online appendix. Because people's problem-solving strategies may rely on an external search,[17] participants were randomly assigned (1:1) to a message informing them that they were allowed versus not allowed to use external resources (Figure 1). The survey was open from 21 February 2013 until 13 May 2013; the survey was closed after 3 reminders were sent to all potential participants, which resulted in the desired sample size approved by the institutional review board (N = 300). The survey was administered using Qualtrics survey software. The study was approved by the University of South Florida Institutional Review Board (No. 9047).

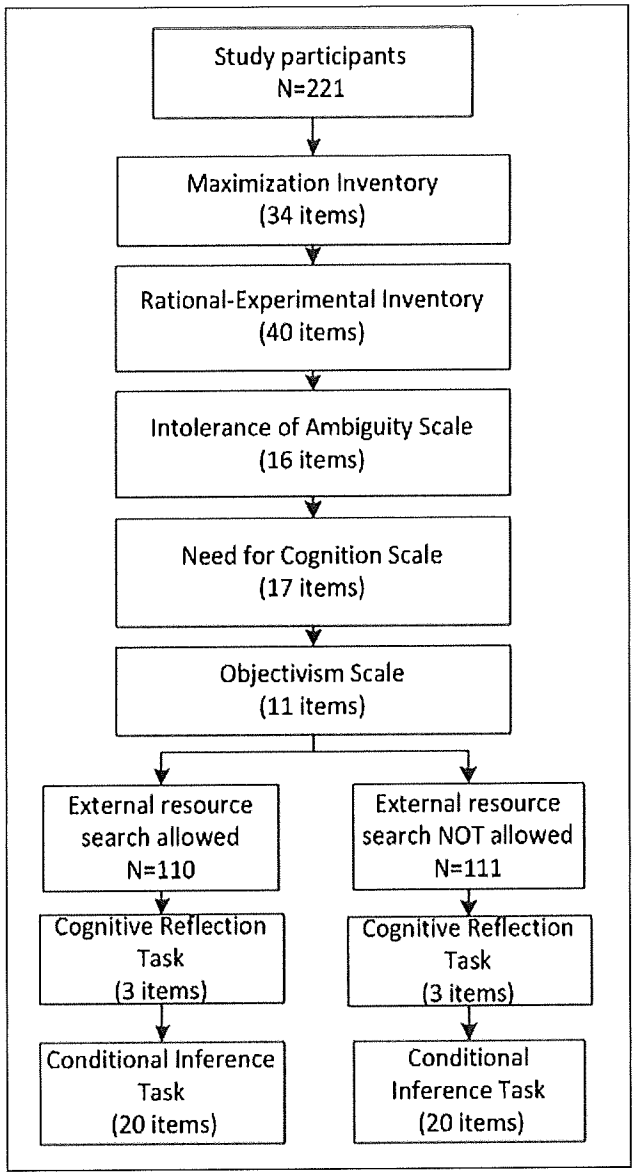


Figure 1 The study design. Because of some evidence that people's problem-solving strategies may rely on the external search, we randomized participants into those who were allowed to use external resources versus those who were not. However, because we found no significant differences in conditional inference between participants with versus without access to external sources, we pooled data from both groups.

Scales to Measure Individual Differences in Cognitive Styles

Maximization Inventory. The Maximization Inventory consists of 3 separate scales.[18] The Alternative Search Scale assesses the tendency to expend resources in exploring all possible opportunities. The Decision Difficulty Scale represents the degree of difficulty experienced when making choices among abundant options. These scales measure maximizing tendency, whereas the Satisficing Scale captures an independent construct. Therefore, maximizing and satisficing are not mutually exclusive; individuals can use both strategies.

Rational-Experiential Inventory. Theories of dual cognition assume 2 distinguishable cognitive styles: rational and intuitive (see also below). The Rational-Experiential Inventory (REI) consists of 2 item subscales and measures intuitive-experiential and analytical-rational thinking based on cognitive-experiential self-theory.[8]

Intolerance of Ambiguity Scale. The Intolerance of Ambiguity Scale (IAS) measures a person's ability to feel comfortable and accept situations in which variables, alternatives, or outcomes are poorly defined, uncertain, or unclear.[5] IAS is particularly applicable to the field of medicine, where individual differences in tolerance of ambiguity are expected to be relevant.

Need for Cognition Scale. The short form of the Need for Cognition Scale contains items that focus on engagement in and enjoyment of cognitive activities.[4] It is a measure of motivated cognition (e.g., information processing, thinking, and judgment).[1] As such, the scale is relevant to individual differences in rational processing.

Objectivism Scale. People differ not only in their cognitive style but also in the kinds of information upon which they base their decisions and beliefs. An objective individual seeks empirical information and attempts to process it in a rational and logical fashion. The Objectivism Scale measures the tendency to base one's judgments and beliefs on empirical information and rational considerations.[6]

Cognitive Reflection Task. The cognitive reflection task (CRT) was designed to test the participants' ability or disposition to suppress intuitive and spontaneous answers in favor of more reflective and deliberative responses.[7] We adapted the CRT using medical contents but preserving the structure of the items to retain fidelity in testing of domain-specific reasoning.

Assessment of Accuracy of Inferences

Belief bias, a central bias studied in dual processing models, is the tendency to be influenced by prior

belief when drawing deductive inference regardless of logical validity.[12–14,19] That is, past work has found that people will exhibit belief bias if an argument is *believable*, even if it defies logic.[9,12–15,20] To assess belief bias in medical situations, we used a conditional inference task.[15]

Table 1 presents the conditional inference types and examples of clinical scenarios that we presented in this format. Believability was pretested using a different sample drawn from the same population as the main study. A total of 48 inference problems were derived by crossing believability with the 4 types of inference problems: modus ponens (MP), denial of antecedent (DA), affirmation of the consequent (AC), and modus tollens (MT). These were administered in a random order using Qualtrics survey software. The complete survey instrument is included in the online appendix.

Statistical Analysis

Descriptive statistics were used to summarize the characteristics of the participants. Reliability of the data obtained by the various scales was assessed using Cronbach's coefficient α. To address the hypothesized age/expertise effect (see Discussion), our main analysis consisted of Pearson's correlation statistics to address key questions: 1) a relationship between age and maximizing and satisficing (and other scales used in our study), 2) a relationship between maximizing and satisficing with other scales used to measure cognitive style. Because the level of training is a proxy for expertise, we also analyzed data accordingly (physicians in training v. attending physicians). The level of cognitive skill is also expected to depend on the medical disciplines, as some fields are more procedure oriented (e.g., surgical disciplines) while others are more cognitively oriented (e.g., internal medicine disciplines). It is also expected that gender may affect cognitive styles.[21–23] We therefore analyzed data according to these a priori defined subgroups. The reported P values were corrected for multiple comparisons using a Bonferroni adjustment. Chi-square test was used to compare differences in responses between trainees and attendings on the conditional inference task. All analyses were performed using statistical package SPSS and Stata.

RESULTS

An invitation to participate in the study was sent by e-mail to 1023 resident, fellow, and attending physicians. Of the 301 physicians who started the survey, 221 completed all questions. There was no significant difference between those who completed and those who started the survey on the collected variables. The sample consisted of 75% (n = 165) trainees (residents and fellows) and 25% (n = 56) attending physicians. Overall, 120 (54%) were male; females comprised 41% (n = 23) of attending physicians and 47% (n = 78) of physicians in training, respectively. Median age in years was 31 (mean 33.7; range 25–69). There was no difference in age between males (median 31; range 25–69) and females (median 30; range 25–59) (P = 0.336). As expected, attending physicians were older (median 42; range 27–69) than trainees (median 29; range 25–55); ($P < 0.0001$) The participants from surgical disciplines comprised 26% (n = 57) of the sample (see Table 2). A majority of the participants completed the survey in less than 1 hour (median 34 minutes; range 8–249).

The survey flow-diagram is presented in Figure 1. Because there were no significant differences in conditional inference between participants with versus without access to external sources, we present pooled data.

The reliability of the data obtained on all 6 scales is shown in Table 3. For each scale, the values obtained were similar to those reported in studies of nonphysicians. For Maximization Inventory we obtained Cronbach's α of 0.746 for the Satisficing Scale, 0.858 for the Decision Difficulty Scale, and 0.879 for the Alternative Search Scale versus 0.73, 0.85, and 0.83, respectively, reported by Turner and others[18]; for the Rational-Experiential Inventory, Cronbach's α was 0.888 for the Rational Scale and 0.893 for the Experiential Scale versus 0.90 and 0.87, respectively, described by Pacini and Epstein[8]; our Cronbach's α for Need for Cognition Scale was 0.894 versus 0.90 reported by Cacioppo and others[4]; for the Objectivism Scale we obtained Cronbach's α of 0.717 versus 0.83 calculated in the original report.[6]

Cronbach's α remained modest for Intolerance of Ambiguity (0.643) although it was very close to the values observed in the other studies reported in the literature (0.64 and 0.63 calculated by Sobal and DeForge[24]). Similarly, Cronbach's α for Cognitive Reflection Task was 0.599. A recent study found Cronbach's α for expanded CRT of 0.67. However, the authors also cautioned that this scale has only 3 items—too few for testing reliability.[25] Nevertheless, the overall results provided sufficient reassurance that we could continue with the analysis as originally planned and that scales developed in nonmedical fields are applicable to physicians (see also limitations in the Discussion section).

Table 1 The Four Inferences Studied Using the Conditional Inference Model with Examples of Invalid, Believable and Invalid, Unbelievable Clinical Scenarios

Inference	Clinical Scenario Example	Major Premise	Minor Premise	Conclusion	Validity
Modus ponens (MP)	Unbelievable Example Assume the following is true: *If a patient has a high fever, then the patient has malaria.* Given that the following premise is also true: *Ms. Boyle has a high fever.* Is it necessary that: *Ms. Boyle has malaria.* ○ Yes ○ No	If A then B	A	B	Valid
Denial of antecedent (DA)	Believable Example Assume the following is true: *If a patient has pulmonary embolism, then the patient is short of breath.* Given that the following premise is also true: *Mrs. Smith does not have pulmonary embolism.* Is it necessary that: *Mrs. Smith is not short of breath.* ○ Yes ○ No	If A then B	Not A	Not B	Invalid
Affirmation of the consequent (AC)	Believable Example Assume the following is true: *If a patient has pulmonary embolism, then the patient is short of breath.* Given that the following premise is also true: *Mrs. Smith is short of breath.* Is it necessary that: *Mrs. Smith has pulmonary embolism.* ○ Yes ○ No	If A then B	B	A	Invalid
Modus tollens (MT)	Unbelievable Example Assume the following is true: *If a patient has a high fever, then the patient has malaria.* Given that the following premise is also true: *Ms. Boyle does not have malaria.* Is it necessary that: *Ms. Boyle does not have a high fever.* ○ Yes ○ No	If A then B	Not B	Not A	Valid

Note: To evaluate the believability of each statement, we first conducted a pilot study. Forty-one participants rated the probability that each of the 20 statements is true, on a scale of 0%–100%. The 6 statements that received the highest median ratings were classified as *believable* and the 6 statements that received the lowest median were classified as *unbelievable* (see Supplementary material at http://mdm.sagepub.com/supplemental for other examples).

Table 2 Characteristics of Participants ($N = 221$)

Variable	No. (%)
Training status	
Trainees (resident/fellow)	165 (75)
Faculty (attending)	56 (25)
Gender	
Male	120 (54)
Female	101 (46)
Median age, years (range)	31 (25–69)
Discipline	
Internal medicine	37 (17)
Pediatrics	29 (13)
Surgery	19 (9)
Obstetrics and gynecology	15 (7)
Radiology	15 (7)
Ophthalmology	12 (5)
Psychiatry	12 (5)
Other	82 (37)
Discipline type	
Surgical	57 (26)
Nonsurgical	164 (74)

Table 3 shows correlation analyses between maximizing-satisficing and other measures of cognitive styles. The tendency to engage in analytical thinking correlated positively with satisficing ($r = 0.226$; $P = 0.032$) and need for cognition ($r = 0.745$; $P = 0.000$). Similarly, objectivism correlated positively with analytic thinking ($r = 0.535$; $P = 0.000$), alternative search ($r = 0.278$; $P = 0.0012$), and need for cognition ($r = 0.358$; $P = 0.000$). Ambiguity intolerance correlated negatively with the need for cognition ($r = -0.528$; $P = 0.000$) and analytical thinking ($r = -0.346$; $P = 0.00$), which correlated negatively with decision difficulty ($r = -0.233$; $P = 0.021$). The latter findings indicate, surprisingly, that when one is faced with a large number of decisions, the higher uncertainty and decision difficulty are associated with reduced application of the rational analytical process. Of note, none of the satisficing-maximizing subscales significantly correlated with the CRT.

Figure 2 shows a correlation between age and maximizing and satisficing. Surprisingly, we found statistically significant negative correlation between age and satisficing ($r = -0.239$; $P = 0.017$). As expected, alternative search decreased with age ($r = -0.220$; $P = 0.047$) but not with dealing with decision difficulty ($r = -0.170$; $P = 0.53$). The Alternative Search Scale showed statistically significant correlation with the Decision Difficulty Scale ($r = 0.415$; $P = 0.000$). On other hand, satisficing displayed no statistically significant correlation with alternative search ($r = 0.185$; $P = 0.32$) or decision difficulty ($r = 0.0421$; $P = 1.00$), findings consistent with the original report of maximizing-satisficing scale used in our study.[18] Age showed no statistically significant correlation with any other scale used in our study for any subgroup analyses.

The subgroup analyses confirmed a positive correlation between analytical thinking and satisficing ($r = 0.265$; $P = 0.025$) and negative correlation between analytical thinking and decision difficulty ($r = -0.264$; $P = 0.028$) in the trainees; the latter indicates that analytical thinking tends to decrease when one is faced with the larger number of decisions. Interestingly, no such a correlation was observed among attendings; here, we detected a positive correlation between satisficing and intuitive-experiential thinking ($r = 0.427$; $P = 0.0369$) but not in the trainees ($r = 0.127$; $P = 1.0$). The correlation coefficients were statistically significantly different between the two subgroups ($P = 0.038$). While we expected a correlation between intuitive-experiential thinking and age, the failure to observe a significant correlation may have been a statistical artifact due to data concentration within relatively a narrow group of the participants. Males showed a positive correlation between alternative search and intolerance toward ambiguity ($r = 0.299$; $P = 0.039$), while no such correlation was observed in the females ($r = 0.0727$; $P = 1.00$). However, the difference between the two correlation coefficients was not significant ($P = 0.084$). We also observed a positive correlation between satisficing and alternative search in males ($r = 0.416$; $P = 0.0001$) but not in females ($r = -0.149$; $P = 1.0$); the difference between two correlation coefficients was highly significant ($P < 0.0001$). Finally, we observed somewhat different patterns of correlation between surgical and nonsurgical disciplines. Surgical participants displayed a positive correlation between analytical thinking and satisficing ($r = 0.468$; $P = 0.009$), while the participants from nonsurgical disciplines displayed a positive correlation between alternative search and decision difficulties ($r = 0.421$; $P = 0.000$). The correlation coefficients between satisficing and analytical thinking were statistically significantly different ($P = 0.029$) between surgical versus nonsurgical participants while no such difference was detected between decision difficulty and alternative search ($P = 0.826$). Both groups showed positive correlation between analytical thinking and need for cognition and objectivism.

CRT score positively correlated with MP inference ($r = 0.203$; $P = 0.0237$) and negatively correlated with the fallacious DA ($r = -0.192$; $P = 0.041$) and AC ($r = -0.306$; $P = 0.000$) inferences but not with MT.

Table 3 Means, Standard Deviations (s), Reliabilities, and Intercorrelations of the Scales Measuring Individual Differences in Cognitive Styles ($N = 221$)

Scale	Mean	s	1	2	3	4	5	6	7	8	9
1. MI: Decision Difficulty	3.200	0.758	**0.858**								
2. MI: Alternative Search	3.925	0.821	0.415[a]	**0.879**							
3. MI: Satisficing	4.860	0.490	0.042	0.180	**0.746**						
4. REI: Rational	2.980	0.531	−0.233[a]	0.021	0.226[b]	**0.888**					
5. REI: Experiential	2.294	0.577	−0.070	0.105	0.198	0.132	**0.893**				
6. Intolerance of Ambiguity	3.068	0.480	0.198	0.194	−0.216[b]	−0.346[a]	−0.141	**0.643**			
7. Need for Cognition	4.241	0.695	−0.172	0.007	0.154	0.745[a]	0.145	−0.528[a]	**0.894**		
8. Objectivism	2.766	0.492	−0.076	0.279[a]	0.154	0.535[a]	−0.081	−0.020	0.358[a]	**0.717**	
9. Cognitive Reflection Task	1.490	1.003	−0.091	−0.088	0.080	0.104	0.042	−0.115	0.107	0.006	**0.599**

Note: MI = Maximization Inventory; REI = Rational-Experiential Inventory. Boldface numbers are scale reliabilities (Cronbach's α). Scale dimensions (higher numbers indicate more of an attribute): MI: Decision Difficulty (1–6); MI: Alternative Search (1–6); MI: Satisficing (1–6); REI: Rational (0–4); REI: Experiential (0–4); Intolerance of Ambiguity (1–6); Need for Cognition (1–6); Objectivism (1–5); Cognitive Reflection Task (0–3).
a. Correlation is significant at the 0.01 level (2-tailed).
b. Correlation is significant at the 0.05 level (2-tailed).

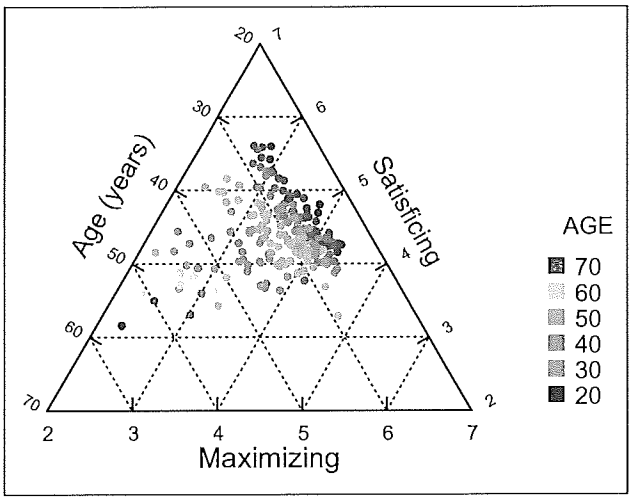


Figure 2 Relationships among age and maximizing and satisficing. As age increased in the physicians sample, problem-solving strategy relied less on satisficing and maximizing (tendency to engage in alternative search to find the optimal solution).

Analytic thinking correlated negatively with the number of incorrect answers on DA ($r = -0.232$; $P = 0.031$), while satisficing correlated negatively with the number of incorrect inferences on AC ($r = -0.25$; $P = 0.02$), both fallacious. Taken together, the results indicate that those individuals who tend to engage in some form of effortful reasoning (measured as cognitive reflection, satisficing, or rational thinking) score better on conditional inference tasks. Despite the positive correlation of trainees with analytical thinking and attendings with intuitive thinking, trainees performed worse on the conditional inference task, endorsing more fallacious AC inferences (25.4% v. 18.8% agreement, $P = 0.0005$; Figure 3), indicating that experience is also important in formal inferential process. In the attempt to delineate the effect of age from experience, we regressed AC inferences on both variables, but this resulted in negative suppression.[26] This occurred because in medicine, as in many professions, age and experience are positively correlated across individuals. (see Discussion)

DISCUSSION

We report the first multidimensional assessment of cognitive styles in physicians with a focus on satisficing and maximizing—arguably one of the key cognitive strategies that humans use. The 2 prior studies of analytical styles among physicians used only one of the instruments, the Rational-Experiential Inventory, used in our study.[27,28]

Several immediate implications emerge from our findings. First is the manner in which our results integrate within current theories of human cognition and whether these findings have implications for revising some of the established concepts related to cognitive style and decision making. Second is the implication of our findings for understanding the relationship between experience (age) and formal training. Third is the extent to which certain cognitive styles may lead to better problem solving and decision making (and hence, by extension, to better patient outcomes), suggesting possible educational and prescriptive remedies that may improve the way the physicians make decisions.

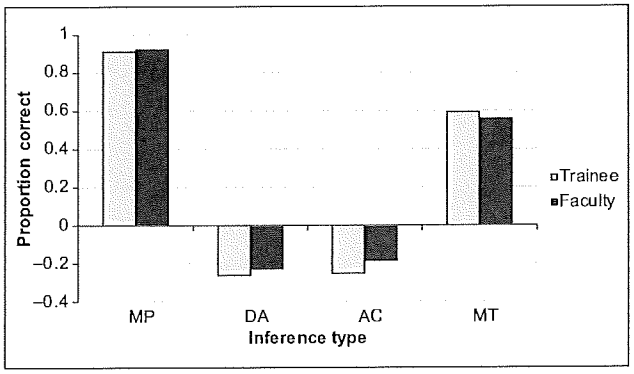


Figure 3 Assessment of agreement on 4 different conditional inference tasks. Note that trainees performed worse in avoiding inferential error related to fallacy of affirming the consequent (AC), while no differences were seen on other syllogism tasks: MP, modus ponens; MT, modus tollens; DA, fallacy of denying the antecedent. Note that the numbers above the x-axis represent the proportion of the participants answering correctly (MP and MT), while the numbers below the x-axis represent the proportion of participants answering incorrectly (DA and AC).

To address the first question, we think that our results can be best explained within a framework of dual processing theories. Dual processing theories postulate that cognition is governed by Type 1 processes (which are intuitive, automatic, fast, narrative, experiential, and affect-based) and Type 2 processes (which are analytical, slow, verbal, and deliberative and support formal logical and probabilistic analyses).[2,10,13,29–36] The current view of leading theorists is that humans are "cognitive misers" with a tendency to use the least possible effort to engage in problem solving and decision making.[10] Type 1 processes do not require working memory, whereas Type 2 processes rely heavily on working memory.[9] A decision maker's initial, default response to the problem is governed by rapid, autonomous, Type 1 processes, which are generated by series of associations ("what first comes to mind"). However, this initial response may or may not be adaptive to reality. To decouple the real world from imaginary situations, decision makers need the capacity to think hypothetically.[9,10] This is provided by Type 2, higher order reasoning and analytical processes, which can interrupt (override) autonomous Type 1 responses.[9,10] The initiation of the override of Type 1 processes by Type 2 processes is different from the actual algorithm of information processing; it depends on the higher level cognitive dispositional functions of reflective mind, which is related to rationality but is distinct from intelligence.[10]

The response to the task at hand starts with serial associative thoughts of Type 1 process, which tend to focus on the single, most relevant option ("relevance and singularity principle").[12,13] If the proposed solution does not "feel right," Type 2 processes intervene.[37] This may take a form of satisficing (carried by the algorithmic mind) or a search for the best possible solution (maximizing).[10] Medical education typically relies on heuristics derived from intensive practice of problem solving and decision making using case vignettes (real or hypothetical), which, after many hours of training, become internalized and represent the basis for the actual practice of medicine. In this context, it is important that both Type 1 and Type 2 processing can generate biases or produce normatively correct answers.[9]

From this perspective, satisficing and maximizing represent 2 distinct forms of decision-making strategies.[18] That is, maximizing is not an inverse of satisficing, as traditionally assumed (and incorrectly measured).[38,39] This can then explain a negative correlation with age and maximizing and satisficing: As they advance in their careers, physicians are less willing to spend time and effort in finding an adequate solution to the clinical problem, whether based on an exhaustive search for best possible or good enough solutions. As they age, physicians tend to move from analytical to intuitive-experiential mode of thinking. Surprisingly, and despite the fact that the concept was introduced more than 5 decades ago,[3,40] we could not find any empirical data evaluating age in relation to maximizing-satisficing. The only study that described cognitive styles related to aging did not measure satisficing directly but expressed it as an inverse of maximizing,[41] the concept that was refuted by Turner and others.[18] As shown in Figure 2, age is independently correlated with satisficing and maximizing. The effect of age is indirectly corroborated by findings that trainees rely more on analytical styles while attendings (typically, older physicians) are inclined to use intuitive-experiential cognitive styles. Nevertheless, these unexpected findings raise questions regarding reinterpretation of satisficing as traditionally understood and point to the need for future replications of our results.

Maximizing and satisficing, therefore, represent 2 different problem-solving strategies. Evidence that satisficing constitutes an effective problem-solving strategy is also reflected by the facts that 1) satisficing correlates positively with disposition to think analytically, which, in turn, resulted in fewer incorrect answers on syllogism tasks, and 2) those participants who scored higher on satisficing dimension had fewer incorrect inferences on the fallacy of affirming

the consequent task. In general, our results show that the tendency to engage in some form of effortful reasoning (as measured by cognitive reflection, satisficing, or rational thinking) positively predicts performance on conditional inference tasks.

Our study sheds some insights into a complex relationship between clinical experience and the effect of formal training on cognitive performance.[42] Some studies suggest that patients' outcomes may be better when patients are treated by younger physicians, who also tend to perform better on board examinations.[43] This could be due to their inclination to rely on analytical reasoning and adhere to guideline-concordant practice,[28] which is also stressed in the courses on evidence-based medicine (EBM) to which many older physicians have not been exposed. Indeed, our findings indicate that physicians who were more advanced in their careers were less willing to spend time and effort in an exhaustive search for best solutions. This appears to point to the obvious educational prescription: Development of evidence-based resources that are user-friendly, reliable, and immediately accessible may enable all physicians—younger and older alike—access to the information at the time they need it. These educational prescriptions may further need to be customized based on the individual's cognitive style. For example, the physicians who display intuitive-experiential cognitive style (attendings in our study) may prefer a narrative, storytelling (case-based) approach, while those who are inclined toward rational thinking (trainees in our study) may prefer EBM, rule-based education with its emphasis on quantitative evidence and guidelines.

However, the situation between experience and expertise is more complex than this discussion appears to indicate. While formal training and continuing medical education are undoubtedly important, so is experience.[44] Because it takes a long time—at least 10 years or 10,000 hours[45]—to acquire important skills, it is typical that older individuals will possess required skills for optimal practice of medicine. Experts' judgments are particularly accurate in the environment that is sufficiently regular to be predictable and following sufficient opportunities to learn these regularities.[30,44] These characteristics are typical of the practice of medicine. Expertise provides individuals with the necessary "mindware."[10] The concept of mindware can explain why older doctors fared better on at least of one formal inferential task despite the fact that the trainees showed a tendency for analytical thinking. It is not enough to be motivated to engage in analytic thinking—one must also have the necessary cognitive tools. Lack of required tools, or existing inappropriate cognitive tools, can lead to erroneous responses even with analytic reasoning. We surmise that being in a training situation motivates trainees to engage in analytic thinking, but, as their training is not yet complete, they have not yet acquired all the necessary cognitive tools to generate normatively correct responses. They think harder but less effectively. This, too, underlines the importance of training and experience to sound medical decision making. Nevertheless, age probably has an independent effect on cognitive styles that is unrelated to acquiring expertise; aside from the possibility of cognitive decline associated with aging, older physicians tend to experience a large decrease in their theoretical knowledge base[43] and may need to develop compensatory mechanisms that tap into their subject expertise.[42] As demonstrated in our study, these mechanisms appear to serve them rather well, helping them to maintain their mindware for effective problem solving. Unfortunately, in our attempt to delineate the effect of age from experience in the conditional inference task, we detected a spurious effect of negative suppression in regression analysis.[26] This occurred because in medicine, as in many professions, age and experience are positively correlated across individuals. This makes it difficult to isolate the unique influence of one (or the other) variable on some third variable (see Beckstead[26] for details). Thus, while we found reliable and systematic differences among physicians as a function of both their age and years of experience, we are not able to make a statement about the relative importance of the cumulative experiences specific to the practice of medicine, and the general biosocial process of aging, as determinants of these differences.

The main limitation of our study is that it was done at a single institution. However, we believe that the overall results are likely generalizable to all U.S. physicians because the composition of physicians at the University of South Florida is similar to what would be expected around the United States. Nevertheless, we cannot rule out the possibility of self-selection of faculty versus trainees, which may have generated an analytical sample producing a potential artifact in our statistical analysis. This possibility can only be confirmed or refuted by future studies aiming to reproduce our research results. In addition, our study represents a cross-sectional look at a point in time of a dynamic phenomenon of the downstream effects of the selection and education of physicians. That is, a proportion of physicians with given expertise and training level may differ if another snapshot

in time is taken. This may explain, for example, why we had few family practice physicians. Nevertheless, we had a wide representation of physicians across the specialties that use similar reasoning skills as family practitioners (e.g., internal medicine, pediatrics, psychiatry, etc), providing sufficient generalizability of our findings. In addition, the results agree well with the findings from the general population, alleviating potential concerns of self-selection.[15]

We also assume that better decision making leads to better patient outcomes, but we have not actually measured the outcomes. Nevertheless, this is an acceptable normative position to take, as previous studies showed that physicians' reactions to ambiguity affect physicians' resource use and practice patterns.[36,46–48] In fact, it is quite likely that cognitive processes play one of the key roles in the wide variation in contemporary medical practice,[49] much of which results in inappropriate diagnostic and treatment decisions. It is estimated that more than 30% of medical interventions are currently not appropriately applied.[50] This implies that many of the features we measured in this study may be immutable. However, it is possible to train people to face uncertainty[36] and to "stop and reflect," which may increase their capacity to resist acting on "what first comes to mind."[30,35,51] This may improve accuracy in inferences and, in turn, lead to better patient outcomes.[35]

FUTURE RESEARCH

Our study lends itself to straightforward extension of this line of research to at least 2 settings: 1) evaluation of cognitive styles of physicians in actual practice and 2) education of medical doctors. For example, it would be interesting to correlate the cognitive scale scores with the patterns of diagnostic testing and treatment prescribing. This would provide direct empirical evidence regarding potential causes of over- and undertesting and over- and undertreatment—a significant public policy question, as discussed above. Similarly, it would be interesting to test whether adaptation of educational practice to cognitive styles improves trainees' knowledge. For example, it would be possible to conduct a randomized trial attempting to answer whether physicians with a tendency toward intuitive-experiential cognitive style score better on medical knowledge tests when the material is presented in a narrative versus algorithmic, rule-based approach. A number of similar proposals can emerge from our current results along the lines highlighted here. In addition, other researchers may find it useful to use our approach, as the appendix provides a ready-to-use protocol, saving time and effort in identifying the most suitable instruments to test the constructs presented in this paper.

In conclusion, we present the first multidimensional study of cognitive styles of physicians, which calls for reinterpretation of some of our prevailing views on human cognition while highlighting additional implications for medical practice and training. Our study also supports dual processing theories underpinning physicians' decision making, which should be taken into account in the ways we train doctors and deliver continuing medical education.

ACKNOWLEDGMENTS

We thank Dr. Pat Croskerry for his input and helpful suggestions related to the earlier version of the paper. Drs. Benjamin Djulbegovic and Jason Beckstead had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. This study was approved by University of South Florida Institutional Review Board No. 9047.

REFERENCES

1. Appelt KC, Milch KF, Handgraaf MJJ, Weber EU. The decision making individual differences inventory and guidelines for the study of individual differences in judgment and decision-making research. Judgm Decis Mak. 2011;6(3):252–62.

2. Stanovich KE, West RF. Individual differences in reasoning: implications for the rationality debate? Behav Brain Sci. 2000;23:645–726.

3. Simon HA. A behavioral model of rational choice. Q J Econ. 1955;69(1):99–118.

4. Cacioppo JT, Petty RE, Kao CF. The efficient assessment of need for cognition. J Pers Assess. 1984;48(3):306–7.

5. Budner S. Intolerance of ambiguity as a personality variable. J Pers. 1962;30:29–50.

6. Leary MR, Shepperd JA, McNeil MS, Jenkins TB, Barnes BD. Objectivism in information utilization: theory and measurement. J Pers Assess. 1986;50(1):32–43.

7. Frederick S. Cognitive reflection and decision making. J Econ Perspect. 2005;19(4):25–42.

8. Pacini R, Epstein S. The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. J Pers Soc Psychol. 1999;76:972–87.

9. Evans JSBT, Stanovich KE. Dual-process theories of higher cognition: advancing the debate. Perspect Psychol Sci. 2013;8(3):223–41.

10. Stanovich KE. Rationality and the Reflective Mind. Oxford (UK): Oxford University Press; 2011.

11. Kruglanski AW, Gigerenzer G. Intuitive and deliberate judgments are based on common principles. Psychol Rev. 2011; 118(1):97–109.

12. Evans JSTBT. The heuristic-analytic theory of reasoning: extension and evaluation. Psychonom Bull Rev. 2006;13:378–95.

13. Evans JSTBT. Hypothethical Thinking. Dual Processes in Reasoning and Judgement. New York (NY): Psychology Press, Taylor and Francis Group; 2007.

14. Evans JSTBT, Curtis-Holmes J. Rapid responding increases belief bias: evidence for the dual theory of reasoning. Think Reason. 2005;11:382–9.

15. Evans JS, Handley SJ, Neilens H, Over D. The influence of cognitive ability and instructional set on causal conditional inference. Q J Exp Psychol. 2010;63(5):892–909.

16. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci. 2010;33(2–3):61–83.

17. Sparrow B, Liu J, Wegner DM. Google effects on memory: cognitive consequences of having information at our fingertips. Science. 2011;333(6043):776–8.

18. Turner BM, Rim HB, Betz NE, Nygren TE. The maximization inventory. Judgm Decis Mak. 2012;7(1):48–60.

19. Evans JSTBT. Dual-process theories of reasoning: contemporary issues and developmental applications. Dev Rev. 2011;31: 86–102.

20. Evans JSBT, Barston JL, Pollard P. On the conflict between logic and belief in syllogistic reasoning. Mem Cogn. 1983;11: 295–306.

21. Byrnes JP, Miller DC, Schafer WD. Gender differences in risk taking: a meta-analysis. Psychol Bull. 1999;125(3):367–83.

22. Croson R, Gneezy U. Gender differences in preferences. J Econ Lit. 2009;47(2):448–74.

23. Gneezy U, Niederle M, Rustichini A. Performance in competitive environments: gender differences. Q J Econ. 2003;118(3): 1049–74.

24. Sobal J, DeForge BR. Reliability of Budner's intolerance of ambiguity scale in medical students. Psychol Rep. 1992;71(1): 15–8.

25. Toplak ME, West RF, Stanovich KE. Assessing miserly information processing: an expansion of the cognitive reflection test. Think Reason. Forthcoming.

26. Beckstead JW. Isolating and examining sources of suppression and multicollinearity in multiple linear regression. Multivariate Behav Res. 2012;47(2):224–46.

27. Calder LA, Forster AJ, Stiell IG, et al. Experiential and rational decision-making: a survey to determine how emergency physicians make clinical decisions. Emerg Med J. 2012;29(10):811–6.

28. Sladek RM, Bond MJ, Huynh LT, Chew DP, Phillips PA. Thinking styles and doctors' knowledge and behaviours relating to acute coronary syndromes guidelines. Implement Sci. 2008;3:23.

29. Kahneman D. Maps of bounded rationality: psychology for behavioral economics. Am Econ Rev. 2003;93:1449–75.

30. Kahnemen D. Thinking Fast and Slow. New York (NY): Farrar, Straus and Giroux; 2011.

31. Slovic P, Finucane ML, Peters E, MacGregor DG. Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. Risk Anal. 2004;24(2):311–22.

32. Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Adv Health Sci Educ Theory Pract. 2009;14(Suppl 1):27–35.

33. Croskerry P. A universal model of diagnostic reasoning. Acad Med. 2009;84(8):1022–8.

34. Croskerry P, Abbass A, Wu AW. Emotional influences in patient safety. J Patient Saf. 2010;6(4):199–205.

35. Croskerry P, Nimmo GR. Better clinical decision making and reducing diagnostic error. J R Coll Physicians Edinb. 2011;41(2): 155–62.

36. Djulbegovic B, Hozo I, Beckstead J, Tsalatsanis A, Pauker SG. Dual processing model of medical decision-making. BMC Med Inform Decis Mak. 2012;12(1):94.

37. Thompson VA, ProwseTurner JA, Pennycook G. Intuition, reason, and metacognition. Cogn Psychol. 2011;63:107–40.

38. Schwartz B, Ward A, Monterosso J, Lyubomirsky S, White K, Lehman DR. Maximizing versus satisficing: happiness is a matter of choice. J Pers Soc Psychol. 2002;83(5):1178–97.

39. Nenkov GY, Morrin M, Ward A, Schwartz B, Hulland J. A short form of the maximization scale: factor structure, reliability and validity studies. Judgm Decis Mak. 2008;3:371–88.

40. Simon H. Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior. New York (NY): Wiley; 1957.

41. Tanius BE, Wood S, Hanoch Y, Rice T. Aging and choice: applications to Medicare Part D. Judgm Decis Mak. 2009;4(1): 92–101.

42. Norman GR, Eva KW. Does clinical experience make up for failure to keep up to date? ACP J Club. 2005;142(3):A8–A9.

43. Choudhry NK, Fletcher RH, Soumerai SB. Systematic review: the relationship between clinical experience and quality of health care. Ann Intern Med. 2005;142(4):260–73.

44. Klein GA. Sources of Power: How People Make Decisions. Cambridge (MA): MIT Press; 1998.

45. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med. 2004;79(Suppl 10):S70–81.

46. Gerrity MS, DeVellis RF, Earp JA. Physicians' reactions to uncertainty in patient care: a new measure and new insights. Med Care. 1990;28(8):724–36.

47. Kassirer JP. Our stubborn quest for diagnostic certainty: a cause of excessive testing. N Engl J Med. 1989;320(22):1489–91.

48. McNeil BJ. Shattuck lecture—hidden barriers to improvement in the quality of care. N Engl J Med. 2001;345(22):1612–20.

49. Song Y, Skinner J, Bynum J, Sutherland J, Wennberg JE, Fisher ES. Regional variations in diagnostic practices. N Engl J Med. 2010; 363(1):45–53.

50. Berwick DM, Hackbarth AD. Eliminating waste in US health care. J Am Med Assoc. 2012;307(14):1513–6.

51. Croskerry P. From mindless to mindful practice—cognitive bias and clinical decision making. N Engl J Med. 2013;368(26): 2445–8.

Identifying homogenous subgroups for individual patient meta-analysis based on Rough set theory

Eleazar Gil-Herrera, Athanasios Tsalatsanis, Ambuj Kumar, Rahul Mhaskar, Branko Miladinovic, Ali Yalcin, and Benjamin Djulbegovic

Abstract— **Failure to detect and manage heterogeneity between clinical trials included in meta-analysis may lead to misinterpretation of summary effect estimates. This may ultimately compromise the validity of the results of the meta-analysis. Typically, when heterogeneity between trials is detected, researchers use sensitivity or subgroup analysis to manage it. However, both methods fail to explain why heterogeneity existed in the first place. Here we propose a novel methodology that relies on Rough Set Theory (RST) to detect, explain, and manage the sources of heterogeneity applicable to meta-analysis performed on individual patient data (IPD). The method exploits the RST relations of discernibility and indiscernibility to create homogeneous groups of patients. We applied our methodology on a dataset of 1,111 patients enrolled in 9 randomized controlled trials studying the effect of two transplantation procedures in the management of hematologic malignancies. Our method was able to create three subgroups of patients with remarkably low statistical heterogeneity values (16.8%, 0% and 0% respectively). The proposed methodology has the potential to automatize and standardize the process of detecting and managing heterogeneity in IPD meta-analysis. Future work involves investigating the applications of the proposed methodology in analyzing treatment effects in patients belonging to different risk groups, which will ultimately assist in personalized healthcare decision making.**

I. Introduction

In medical research, meta-analysis is used to obtain pooled estimates of the treatment effects reported in various clinical research studies. The importance of meta-analysis stems from the necessity to combine research findings that if considered separately they would produce insignificant, non-generalizable, and unavailing results, unfit to inform medical practice. By systematically combining findings from similar studies it is possible to achieve the totality of evidence necessary to evaluate the efficacy of an investigated treatment.

The challenge researchers face when performing meta-analysis is how to integrate studies that present differences in the design, characteristics and reported effects. Such differences are formally acknowledged as heterogeneity and they are defined as any kind of variability among studies [1]. Typically, there are three types of heterogeneity found in meta-analyses: 1. *Methodological*, which refers to variability in the study design and risk of bias (e.g. randomization, allocation concealment, blindness etc.)[2, 3], 2. *Clinical*, which refers to the variability in the participants, interventions and outcomes studied (e.g. age, race, disease severity, disease progression, past treatment etc.) [2, 3], and 3. *Statistical*, which refers to variability in the observed outcomes [1, 3]. Failure to detect heterogeneity leads to misinterpretation of the summary effect estimates, which jeopardize the quality of the meta-analyses [2, 3] and may produce faulty estimations of the effects magnitude [4, 5]. Both methodological and clinical heterogeneity may result in statistical heterogeneity [6]. Researchers focus primarily on detecting statistical heterogeneity and subsequently on determining whether such heterogeneity is caused due to methodological or clinical variations between studies [1].

Assessing statistical heterogeneity relies on approaches that involve hypothesis testing [1, 7-9], such as the Chochrane's chi-square (Q) [10] and the I^2 measure [9, 11]. Higher values on these tests indicate high heterogeneity between studies. Both chi-square and I^2 tests focus on detecting heterogeneity yet are unable to identify the specific causes that underlie heterogeneity across studies [12]. The burden of explaining heterogeneity falls on the researcher.

To explore and explain the observed heterogeneity, meta-analysts conduct sensitivity analysis, based on the methodological quality of studies, and sub group analysis, based on a pre-specified trial or patient characteristics [3]. That is, the trials included in the meta-analysis are grouped according to pre-specified criteria. In case of individual patient data meta-analysis patients are grouped according to pre-specified clinical characteristics. However, these pre-specified criteria and clinical characteristics are generated in an ad-hoc manner and rely on the skills and medical knowledge of the researcher performing the meta-analysis. Thus, the results of meta-analysis may potentially differ depending on the experience of the meta-analyst.

Subgroup analysis [13] and meta-regression[14] are also applied to individual patient datasets (IPD) containing patient characteristics that may potentially influence the treatment effects. Determining which set of characteristics can be used to obtain homogeneous groups yields in a complex process, where subgroup analysis and meta-regression have been found prone to false positive results and ecological bias.

In this paper, we focus on meta-analyses of individual patient data and we propose a novel methodology to identify homogeneous groups of patients for managing the detected

This work was supported by the Department of Army under grant # W81 XWH-09-2-0175. The study was approved by the University of South Florida institutional review board (IRB # 100701).

E. Gil-Herrera A. Tsalatsanis, A. Kumar, R. Mhaskar, B. Miladinovic and B. Djulbegovic are with the Division of Evidence Based Medicine, Dept. of Internal Medicine, University of South Florida, Tampa, FL 33612, USA (e-mails: eleazar, atsalats, akumar1, rmhaskar, bmiladin, bdjulbeg @health.usf.edu)

A. Yalcin is with the Department of Industrial and Management System Engineering, University of South Florida, Tampa, FL 33620, USA (e-mail: ayalcin@usf.edu).

heterogeneity. Our approach is based on Rough Set Theory (RST) [15] and has the potential to automatize the process of creating subgroups of patients with similar characteristics.

The mathematical principles that govern RST rely on the relations between objects. Using RST, we analyze and evaluate all possible relations between patients to obtain the minimum and dispensable information required to generate homogeneous subgroups of patients (i.e. patients with similar characteristics). We envision our methodology to operate in an automatic manner without the researcher intervention in selecting those characteristics that matter in grouping patients for meta-analyses.

II. METHODOLOGY

A. Dataset

Our dataset consists of individual patient data collected from nine randomized trials studying the effect of Allogeneic Peripheral Blood Stem-cell transplantation (PBSCT) compared to Bone Marrow transplantation (BMT) in the management of hematologic malignancies [16]. In total, 1,111 patients were enrolled. Records of 44 patients containing missing information were removed leaving the dataset with 1067 complete cases. Table 1 describes the details of our dataset.

Table 1. Dataset description

Variable	Description	Categories	%
Age	Patient age	0: <20	6.25 %
		1: (20,40]	47.82%
		2: (40, 65]	45.93%
Gender	Patient gender	1: Male	59.66%
		2: Female	40.34%
Diag	Diagnosis category	Acute lymphoblastic leukemia (ALL)	12.5%
		Acute myelogenous leukemia (AML)	33.52%
		Chronic lymphocytic leukemia (CLL)	0.28%
		Chronic myelogenous leukemia (CML)	43.47%
		Hodgkin's disease (HD)	0.09%
		Idiopathic myelofibrosis (IMF)	0.76%
		Myelodysplastic symdrome (MDS)	5.87%
		Multiple myeloma (MM)	1.04%
		Non-hodking lymphoma (NHL)	2.46%
StatTrans	Diagnosis status	0: Favorable (early-stage disease)	74.62%
		1: Unfavorable (late-stage disease)	25.38%
Mtx	Methotrexate for GVHD prophylaxis	1: Yes	43.84%
		0: No	56.15%
CondReg	Conditioning regimen used	1: Total body irradiation based (TBI)	41.19%
		2: Non TBI based	58.81%
GrowthFac	Use of post-transplantation growth factor	1: G-CSF	58.14%
		0: not used	41.85%
Alloc	Treatment	1: PBSCT	49.05%
		2: BMT	50.95%
Trial	Origin of the study	BR	5.30%
		US1	16.29%
		No	5.78%
		SA	5.10%
		FR	9.56%
		EBMT	30.21%
		CAN	20.36%
		US2	1.70%
		UK	3.69%
Death	Overall survival	0: Survive	59.75%
		1: Death	40.25%

B. Rough Set Theory

In RST, a dataset is represented by an information system defined as a pair $S = (U, A)$ where U is a non-empty finite set of objects that in our case represents the 1,111 patients. The set A represents a non-empty finite set of attributes called the condition attributes that corresponds to the characteristics of each patient. For every attribute $a \in A$, the function $U \rightarrow V_a$ makes a correspondence between an object (i.e. a patient) in U to an attribute value, which is called the value set of a. For example, from table 1, the value of the attribute "Age" can be 0, 1 or 2 for a given patient. A dataset including an outcome variable $d \notin A$, is termed as a decision system, defined as: $DS = (U, A \cup \{d\})$. The decision attribute in our data is the variable *"Death"* representing the overall survival of a patient given the characteristics described in A.

C. Indiscernibility and discernibility relations

Two objects (e.g. patients) $u, u' \in U$ are indiscernible with respect to a set of condition attributes $B \subseteq A$ if they have exactly the same values in all attributes, i.e: $a(u) = a(u')\ \forall\ a \in B$. This relation is called *indiscernibility relation* and is defined as:

$$IND(B) = \{(u, u') \in U^2 : \forall a \in B, a(u) = a(u')\} \quad \forall\ B \subseteq A \quad (1)$$

The *indiscernibility relation* captures the redundant information in the dataset. Every subset $B \subseteq A$, can be used for constructing this relation, however, only subsets that maintain the structure of the original dataset, i.e: $IND(B) = IND(A)$, are considered appropriate. Such a subset $B \subseteq A$, is termed as an exact reduct. In the case that it would not be possible to obtain an exact reduct, approximated reducts with acceptable quality of approximation are considered. The quality of approximation (α_B) of a reduct B quantifies the proportion of objects correctly allocated in a decision class by using only the attributes in B, i.e:

$$\alpha_B = \frac{|POS(B)|}{|U|} \quad (2)$$

where, $POS(B)$ is the set of all objects correctly assigned to the right decision class. In general, the higher the value of α_B, the more desirable the reduct is for constructing homogeneous subgroups.

On the other hand, the *discernibility* relation accounts for differences between objects in terms of their attribute values, i.e:

$$DIS_{DS}(B) = \{(u, u') \in U^2 : \exists a \in B, a(u) \neq a(u')\} \quad \forall\ B \in A \quad (3)$$

III. IDENTIFYING HOMOGENEOUS SUBGROUPS IN INDIVIDUAL PATIENT DATASET

We use the indiscernibility relation to build homogenous subgroups based on patients with the same characteristics and we use the discernibility relation to explore the characteristics that differentiate each subgroup. Fig. 1 depicts an overview of the proposed methodology, which is comprised of 4 processes: 1. Obtain reducts; 2. Create homogeneous groups; 3. Regroup based on similarities; and 4. Evaluate groups' heterogeneity.

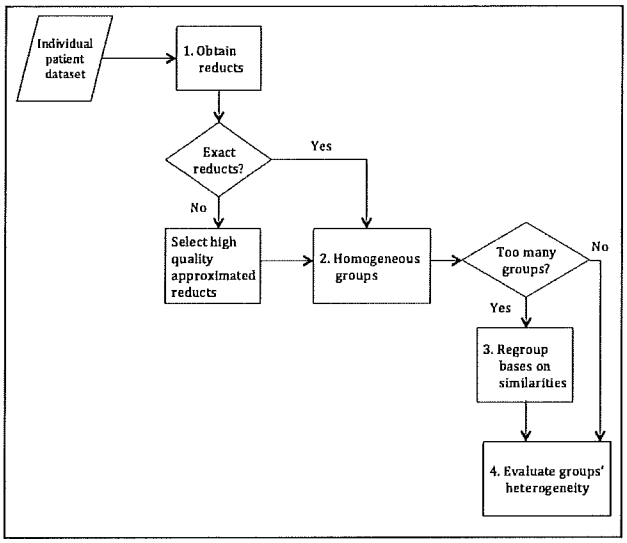


Figure 1. Overview of the RST based methodology for identifying homogeneous subgroups in IPD

Obtaining reducts: First, we use the indiscernibility relation $IND(B)$ to obtain an appropriate subset of condition attributes B as the basis to generate the homogeneous subgroups of patients. To find this subset of attributes (reducts), we use approximated solutions described in [17]. In our dataset, the set $B = \{Age, Diag, StatTrans\}$ stands as the approximated reduct with the highest quality of approximation ($\alpha_B = 0.71$) among all the generated reducts.

Homogeneous groups: The indiscernibility relation partitions the IPD in 32 disjoint homogeneous subgroups with around 40% of them containing less than 10 patients. Subgroups with small number of patients do not include patients from all trials and are unsuitable for an individual patient meta-analysis.

Regrouping process: We obtain subgroups with a larger number of patients by merging smaller subgroups based on a similarity relationship. The similarity relation [18] is defined as a less rigorous version of the indiscernibility relation and is subject to a threshold value that allows small differences considered insignificant. Formally, we define the similarity relation between subgroups as:

$$g_1\, SIM_{B,\gamma}\, g_2 \; iff \; \frac{|X|}{|B|} \geq \gamma, \forall\, g_1, g_2 \in U/IND(B) \; and \; u \in g_1 and \; u' \in g_2 \quad (4)$$

Where, $X = \{a \in B: a(u) = a(u')\}$ and $\gamma \in [0,1]$ is the similarity threshold.

Since comparing all possible combinations between two groups to determine their similarity is a complex process we use a more straightforward procedure consisting in evaluating the differences between subgroups. Then, subgroups having similar differences to the rest of the subgroups are combined resulting in one homogenous group.

We define a discernibility matrix of subgroups $\mathcal{M}_B$, where each cell $\mathcal{M}_B(g_i, g_j)$ represents the number of attributes in B, whose values distinguish subgroup g_i from subgroup g_j, i.e:

$$\mathcal{M}_B(g_i, g_j) = \{|Dif|\}, where\ Dif = \{\, a \in B: a(u) \neq a(u')\} \; \forall\, g_1, g_2 \in U/IND(B) \; and \; u \in g_1 and \; u' \in g_2 \quad (5)$$

Fig. 2 shows a portion of the discernibility matrix obtained for the 32 homogenous subgroups.

IV. RESULTS

The initial 32 homogeneous subgroups are regrouped based on similarities in the number of attributes that distinguish them from the rest of groups. We chose a $\gamma = 0.8$ value (Equation 5) as a threshold parameter of similarity to minimize the number of homogeneous groups by allowing some degree of differences. For example, the initial subgroups 18, 19 and 20 (Fig. 2) can be regrouped since there are no more than 20% of differences across their corresponding rows. In other words, the three subgroups have similar distances, in terms of differences, to the rest of groups. As a result, the 32 homogeneous groups are gathered in three groups. Table 2 shows the homogenous groups resultant after the regrouping process. The mean number of patients in each group is equal to 355 with a standard deviation of 39.15.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2	1																			
3	2	1																		
4	2	1	1																	
5	1	2	2	2																
6	2	1	1	1	1															
7	1	2	2	2	1	2														
8	2	1	1	1	2	1	1													
9	2	1	1	1	2	1	2	1												
10	1	2	3	3	2	3	2	3	3											
11	2	1	2	2	3	2	3	2	2	1										
12	2	3	2	2	2	3	2	3	3	1	2									
13	3	2	1	1	3	2	3	2	2	2	1	1								
14	2	3	3	3	1	2	2	3	3	1	2	1	2							
15	3	2	2	2	2	1	3	2	2	2	1	2	1	1						
16	3	2	2	2	3	2	3	2	2	2	1	2	1	2	1					
17	2	3	3	3	2	3	1	2	3	1	2	1	2	1	2	2				
18	3	2	2	2	3	2	2	1	2	2	1	2	1	2	1	1	1			
19	3	2	2	2	3	2	3	2	2	2	1	2	1	2	1	1	2	1		
20	3	2	2	2	3	2	3	2	1	2	1	2	1	2	1	1	2	1	1	

Figure 2. A portion of the discernibility matrix obtained for the homogeneous groups. Each cell shows the number of attributes that differentiate between each pair of subgroups.

Table 2. Three homogeneous groups obtained from the regrouping process		
Group number	Original group	Number of patients
1	10, 12, 14, 17	392
2	21, 23, 26, 29	359
3	1-9,11,13,15-16,18-20,22,24-25,27-28,30-32	314

The obtained homogeneous groups (Table 2) contain similar distributions in terms of trials, diagnosis and treatment. The statistical heterogeneity (I^2) indicate a negligible heterogeneity value for all the three groups (16.8% in group 1, 0% for group 2, and 0% for group 3), which suggests that all groups are indeed homogeneous.

V. CONCLUSIONS

In this preliminary work, we utilized a methodology typically found in engineering applications to solve a problem that exists in the realm of evidence-based medicine. Researchers who perform evidence synthesis are faced with the challenge of detecting heterogeneity between clinical trials and then explaining it by hypothesizing standards of similarity. However, there is no commonly accepted approach to identify similarities between trials and meta-analysts resolve to ad-hoc solutions. Here we presented a methodology based on Rough Set Theory that has the potential to automatize and standardize this process.

We demonstrated the effectiveness of our methodology using a sample dataset containing 1,111 patients from 9 different trials. We showed that were able to identify the appropriate patient characteristics to construct homogenous groups that presented similar proportion of trials, controls (diagnosis) and interventions (treatments) in accordance to the fundamental doctrine of meta-analysis. Thus, these groups are suitable to derive the pooled estimate of treatment effects in individual patient meta-analysis.

Other applications of this methodology include identifying subgroups of patients that need different treatments, patients with differential responses to therapy, or patients that belong to different risk groups. Analyzing the effect of treatment in each subgroup is very important for personalized healthcare. Our intention is to compare this methodology with similar approaches in other data sets.

Finally, this is a preliminary work and presents limitations. Particularly, we have not investigated the effects of our methodology in the results of meta-analysis, which we intent to do in the future. Other future research includes generalization of our methodology to accommodate clinical trial data in addition to individual patient data.

REFERENCES

[1] J. P. T. Higgins, S. Green, and C. Collaboration, *Cochrane handbook for systematic reviews of interventions* vol. 5: Wiley Online Library, 2008.

[2] S. L. West, G. Gartlehner, A. J. Mansfield, C. Poole, E. Tant, N. Lenfestey, L. J. Lux, J. Amoozegar, S. C. Morton, and T. C. Carey, "Comparative effectiveness review methods: clinical heterogeneity," 2010.

[3] J. Higgins and S. Green. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. Available: http://www.cochrane.org/resources/handbook/index.htm

[4] K. Schulz, "Meta-analyses of interventional trials done in populations with different risks," *Lancet*, vol. 345, p. 1304, 1995.

[5] E. M. Balk, P. A. L. Bonis, H. Moskowitz, C. H. Schmid, J. P. A. Ioannidis, C. Wang, and J. Lau, "Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials," *JAMA: the journal of the American Medical Association*, vol. 287, pp. 2973-2982, 2002.

[6] S. G. Thompson, "Why sources of heterogeneity in meta-analysis should be investigated.," *BMJ*, vol. 19, pp. 1351-1355, 1994.

[7] J. P. Ioannidis, "Interpretation of tests of heterogeneity and bias in meta-analysis," *Jounral of Evaluation in Clinical Practice*, vol. 14, pp. 951-957, 2008.

[8] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, "Identifying and Quantifying Heterogeneity," in *Introduction to Meta-Analysis*, ed: John Wiley & Sons, Ltd, 2009, pp. 107-125.

[9] J. Higgins and S. G. Thompson, "Quantifying heterogeneity in a meta-analysis," *Statistics in medicine*, vol. 21, pp. 1539-1558, 2002.

[10] W. G. Cochran, "Some Methods for Strengthening the Common $\chi 2$ Tests," *Biometrics*, vol. 10, pp. 417-451, 1954.

[11] J. Higgins, S. Thompson, J. Deeks, and D. G. Altman, "Measuring inconsistency in meta-analyses," *BMJ*, pp. 327-357, 2003.

[12] S. G. Thompson and S. J. Sharp, "Explaining heterogeneity in meta-analysis: a comparison of methods," *Statistics in medicine*, vol. 18, pp. 2693-2708, 1999.

[13] S. W. Wang R Fau - Lagakos, J. H. Lagakos Sw Fau - Ware, D. J. Ware Jh Fau - Hunter, J. M. Hunter Dj Fau - Drazen, and J. M. Drazen, "Statistics in medicine--reporting of subgroup analyses in clinical trials.", *New England Journal of Medicine*, vol. 357, pp. 2189 - 2194, 2007.

[14] J. Berlin Ja Fau - Santanna, C. H. Santanna J Fau - Schmid, L. A. Schmid Ch Fau - Szczech, H. I. Szczech La Fau - Feldman, and H. I. Feldman, "Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head.", *Statistics in Medicine*, vol. 21, pp. 371-87, 2002

[15] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*. Norwell, MA, 1992.

[16] H. Florida, "Allogeneic peripheral blood stem-cell compared with bone marrow transplantation in the management of hematologic malignancies: an individual patient data meta-analysis of nine randomized trials," *J Clin Oncol*, vol. 23, pp. 5074-5087, 2005.

[17] S. Vinterbo and A. Øhrn, "Minimal approximate hitting sets and rule templates," *International Journal of Approximate Reasoning*, vol. 25, pp. 123-143, 2000.

[18] R. Slowinski and D. Vanderpooten, "Similarity relation as a basis for rough approximations," 1995.

External Validation of a Web-Based Prognostic Tool for Predicting Survival for Patients in Hospice Care

ARTICLE *in* JOURNAL OF PALLIATIVE CARE · SEPTEMBER 2013

Impact Factor: 0.68 · Source: PubMed

6 AUTHORS, INCLUDING:

Branko Miladinovic

University of South Florida

73 PUBLICATIONS **283** CITATIONS

SEE PROFILE

Rahul Suresh Mhaskar

University of South Florida

75 PUBLICATIONS **359** CITATIONS

SEE PROFILE

Ambuj Kumar

University of South Florida

168 PUBLICATIONS **1,582** CITATIONS

SEE PROFILE

Benjamin Djulbegovic

University of South Florida

364 PUBLICATIONS **12,244** CITATIONS

SEE PROFILE

External Validation of a Web-Based Prognostic Tool for Predicting Survival for Patients in Hospice Care

Branko Miladinovic, Rahul Mhaskar, Ambuj Kumar, Sehwan Kim, Ronald Schonwetter, and Benjamin Djulbegovic

B Miladinovic (corresponding author): Center for Evidence-Based Medicine and Health Outcomes Research, Morsani College of Medicine, University of South Florida, 3515 E. Fletcher Avenue, MDT1200, Mail Code MDC27, Tampa, Florida 33612, USA; bmiladin@health.usf.edu

R Mhaskar, A Kumar: Center for Evidence-Based Medicine and Health Outcomes Research, University of South Florida, Tampa, Florida, USA; **S Kim, R Schonwetter**: HPC Healthcare, Temple Terrace, Florida, USA; **B Djulbegovic**: Center for Evidence-Based Medicine and Health Outcomes Research, University of South Florida, and H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA

Keywords: Prognostat, survival prognostication, Palliative Performance Scale, external validation, Royston-Parmar survival models

Abstract / Prognostat is an interactive Web-based prognostic tool for estimating hospice patient survival based on a patient's Palliative Performance Scale (PPS) score, age, gender, and cancer status. The tool was developed using data from 5,893 palliative care patients, which was collected at the Victoria Hospice in Victoria, British Columbia, Canada, beginning in 1994. This study externally validates Prognostat with a retrospective cohort of 590 hospice patients at LifePath Hospice and Palliative Care in Florida, USA. The criteria used to evaluate the prognostic performance were the Brier score, area under the receiver operating curve, discrimination slope, and Hosmer-Lemeshow goodness-of-fit test. Though the Kaplan-Meier curves show each PPS level to be distinct and significantly different, the findings reveal low agreement between observed survival in our cohort of patients and survival predicted by the prognostic tool. Before developing a new prognostic model, researchers are encouraged to update survival estimates obtained using Prognostat with the information from their cohort of patients. If it is to be useful to patients and clinicians, Prognostat needs to explicitly report patient risk scores and estimates of baseline survival.

Résumé / Le logiciel interactif Prognostat est un instrument de pronostic qui a été créé pour estimer le taux de survie des patients en soins palliatifs. Il est basé sur les résultats de l'indice fonctionnel de l'échelle de rendement palliatif, (ERP) l'âge et le sexe des malades et le stade d'évolution du cancer. Cet outil a été développé à l'aide de données provenant de 5,893 patients hébergés, à partir de 1994, au Victoria Hospice en Colombie Britannique dans l'ouest canadien. Notre étude avait pour but de valider ce logiciel à l'externe avec une cohorte de 590 patients soignés au LifePath Hospice and Palliative Care, Temple Terrace, en Floride aux Etats-Unis. Les critères utilisés pour évaluer l'efficacité de rendement de ce logiciel de pronostic étaient l'indice de Brier, la courbe d'efficacité du récepteur, la discrimination dans la pente d'amplitude, et le test de calibration Hosmer-Lemeshow. Même si les courbes Kaplan-Meier indiquent que chaque niveau de l'échelle de rendement en soins palliatifs est distinct de façon significative, les résultats démontrent peu de similitude entre le taux de survie chez notre cohorte de patients et le taux de survie estimé par le logiciel Prognostat. Avant de procéder à l'élaboration d'un nouveau modèle de prédiction, on encourage les chercheurs à actualiser leurs pronostics de survie obtenu avec Prognostat et à les comparer aux résultats provenant de leur cohorte de patients. Si Prognostat veut devenir un outil utile pour les patients et cliniciens, il devra ouvertement indiquer le score de risques chez les patients et estimer le taux de survie de base.

INTRODUCTION

Accurate prognostication of hospice patient survival gives patients and their family members a vital opportunity to attend to matters such as planning, prioritizing, and preparing for death (1). Predicting patient survival without using a prognostic model is often affected by optimism or avoidance, which can lead to poor prediction of life expectancy. Studies have shown that clinicians consistently overestimate survival times of terminally ill patients (2-4). One prospective cohort study suggested that doctors overestimated survival of terminally ill patients by a factor of 5 (5). Successful prognostication of patient survival depends on developing and testing prognostic models, which entails having accurate patient data for prognosis and selecting clinically relevant candidate predictors and measures of model per-

formance, usually in the context of a multivariable regression survival model (6). This process produces patient performance scores that allow for classification of patients into different risk groups.

The usefulness and validity of a prognostic model are judged by how well the model performs for patients who come from different centres (7). A validated prognostic model is generally accepted to be one that works in a data set other than the one that has been used to develop it (7, 8). There is also a general concurrence that the validation process should follow guidelines and that unvalidated prognostic models should not be applied in clinical practice (9-11). As the value of any prediction model is its generalizability to other groups of patients, our goal was to externally validate Prognostat (12) — a Web-based interactive prognostic tool for estimating hospice patient survival — on a retrospective cohort of 590 hospice patients in Florida, USA. Prognostat estimates survival times based on palliative patients' age group, gender, diagnosis, and score on the Palliative Performance Scale (PPS) (13).

In this paper, we discuss Prognostat and introduce the measures of model performance. Since predictive performance may decrease when Prognostat is tested with new patients as compared to the patients who were used to develop the model, we also discuss a strategy for updating Prognostat in future studies.

METHODS

Study Sample and Survival Estimation Using Prognostat

The patient data were obtained from LifePath Hospice and Palliative Care, licensed since 1983 to serve Hillsborough County, Florida. The data for 590 consecutive deceased patients was extracted starting in January 2009 and working backwards. This study was a retrospective review of deceased patients' medical records, and only data that pertained to outcomes was collected; personal information was not collected, and data were de-identified prior to analysis. A trained nurse assigned PPS scores at admission to our cohort of patients. The University of South Florida's institutional review board approved the study. Two research assistants extracted all data necessary to populate the model variables, and two faculty members (RM and BD) randomly checked 25 percent of the data for accuracy.

Prognostat was developed at the University of Victoria (in Victoria, British Columbia, Canada) using retrospective survival estimates of 5,893 palliative care patients collected at the Victoria Hospice starting in 1994. It calculates survival rate in days for the variables or covariates found to be statistically significant predictors of patient survival — namely, the patient's gender, age group (19 to 44, 45 to 64, 65 to 74, 75 to 84, or 85 and over), diagnosis (lung cancer, breast cancer, colorectal cancer, prostate cancer, other cancer, or noncancer illness), and PPS score.

Decisions regarding hospice admission depend on the care an individual requires and the specific hospice setting. While US Medicare guidelines state that only individuals with a life expectancy of six months or less may be admitted to hospice in the US, the criteria for hospice admission in Canada vary among geographical areas and among individual hospices — that is, some Canadian hospices admit patients with a life expectancy of one month or less, while others do not impose such restrictions. Palliative care providers or programs will often assist patients in determining the best timing for admission to hospice.

The PPS was developed and reported by Anderson and colleagues (13) to measure the functional status of patients receiving palliative care. The scale has 11 possible mutually exclusive levels, from 0 (the patient is dead) to 100 (the patient is ambulatory and healthy). Numerous studies have assessed its performance in a variety of settings and found it to be a statistically significant risk score for calculating survival estimates (14-22).

Prognostat survival estimates were derived using the Cox proportional hazards (CPH) model, which relies on both the baseline survival function and risk scores to estimate patient survival. Because reporting the baseline function under CPH is not possible and Prognostat does not explicitly report prognostic indices (or risk factors), it makes model calibration in other populations unfeasible.[1]

Assessment of Model Performance

Using measures of accuracy, discrimination, and calibration, we analyzed Prognostat's predictive performance based on the ability of the estimated risk score to predict survival. Accuracy refers to the difference between the probability of survival predicted with Prognostat and observed patient survival. The Brier score is a quadratic scoring rule that calculates the differences between actual outcomes and predicted probabilities (23). Given the predicted probability of survival p_i at time t for patient i, and Y_i binary (0-1, dead-alive) variable, the Brier score is defined as: $\sum_i(Y_i\ (1 - p_i)^2 + (1 - Y_i)p_i^2)$. A Brier score of 0 indicates a perfect model, while 0.25 indicates a noninformative model (the value achieved when issuing a predicted probability of 50 percent to each patient). The Brier score may be scaled by its maximum $\text{Brier}_{max}=(1 - \text{mean}(p_i))\ \text{mean}(p_i)$ to obtain Brier_{scaled}

$=\left(1-\frac{\text{Brier}}{\text{Brier}_{\text{max}}}\right)100$ percent, which has interpretation similar to the Pearson correlation coefficient (24).

Calibration refers to how closely the predicted survival calculated at a prespecified time using Prognostat agrees with the observed survival. Since calibration is essentially a test of fit, we applied the Hosmer-Lemeshow (HL) test (25) on the dead-versus-alive binary outcome. The HL chi-squared statistic involves grouping the observations (most commonly in deciles) based on the predicted probabilities and then testing the hypothesis that the difference between observed and predicted events is simultaneously zero for all the groups. This test is equivalent to testing the hypothesis that the observed number of events in each of the groups is equal to the expected number of events based on the fitted model. The higher the HL p-value, the better calibrated the model is. The HL calibration can be visually expressed by plotting deciles of predicted-versus-observed proportions of survival at each time point.

Discrimination is the ability of the model to differentiate between the patients who died versus those who survived at a pre-specified time. A rank order statistic commonly used to summarize discrimination with and without the outcome has been the area under the receiver operating curve (AUC) (26), which is a plot of the sensitivity (true positive rate) against 1-specificity (false positive rate) for consecutive cutoffs of the probability of an outcome. The maximum value of the area under the receiver operating curve (AUC), AUC=1, indicates a perfect prediction model, while a value of AUC=0.5 indicates that 50 percent of patients have been correctly classified (as good as by chance). As a rank order statistic, AUC is insensitive to errors such as difference in average survival. For this reason, a model can have relatively moderate AUC scores and at the same time be inaccurate and have high Brier scores (or low-scaled Brier scores).

The discrimination slope is a measure of how well subjects with and without the outcome are separated. It is defined as the absolute difference in mean predictions of survival (mean $[p_i]$) between those who died and those who survived at time t (8). Because it is an overall measure of differences in mean survival probabilities, in addition to the discrimination slope we have used box plots to assess the extent to which survival differentiation at each time point is achieved for all survival estimates. All statistical calculations were performed using Stata version 11.2.

RESULTS

Patient characteristics of the retrospective cohort are summarized in Table 1. The extracted data were found to be in substantial agreement (kappa=0.85). In addition to presenting data for our cohort of 590 patients, in each column, as a second cell entry, we present data from the Victo-

Table 1 / Patient Characteristics and Survival Times by Age, Gender, Cancer Diagnosis, and PPS*

	Survival times (in days)			
Variable	Mean (95% CI**)	Median (95% CI)	Range	Number of patients
Overall	14 (12, 17)	6 (5, 6), 8 (7.5, 8.5)***	1-371	590, 5,893***
Age at treatment				
<45	15 (8, 22)	8 (4,12)	1-95	37 (6.3%), 245 (4.2%)
45-64	14 (11, 17)	7 (5, 9)	1-114	187 (31.7%), 1,270 (21.6%)
65-74	14 (8, 20)	5 (4, 6)	1-271	110 (18.6%), 1,489 (25.3%)
75-84	14 (8, 20)	6 (5, 7)	1-371	129 (21.9%), 1,945 (33.0%)
85+	15 (9, 21)	5 (4, 6)	1-313	127 (21.5%), 944 (16.0%)
Gender				
Male	14 (10, 18)	6 (5, 7)	1-371	293 (49.7%), 2,822 (48.9%)
Female	15 (11, 19)	6 (5, 7)	1-271	295 (50%), 3,071 (51.1%)
Number of patients with cancer				
Noncancer	12 (8, 16)	5 (4, 6)	1-371	363 (61.5%), 928 (15.7%)
Cancer	17 (14, 20)	9 (7, 11)	1-113	227 (38.5%), 4,965 (84.3%)
Initial PPS score*				
PPS 10%	5 (3, 7)	3 (2, 4), 1 (0.9, 1.1)	1-77	188 (32.6%), 569 (9.6%)
PPS 20%	16 (8, 24)	5 (4, 6), 2 (1.8, 2.2)	1-371	125 (21.7%), 732 (12.5%)
PPS 30%	15 (11, 19)	7 (5,9), 5 (4.6, 5.4)	1-140	123 (21.4%), 1,403 (23.9%)
PPS 40%	24 (18, 30)	14 (11, 17), 13 (11.9, 14.1)	1-174	96 (16.7%), 1,590 (27.0%)
PPS 50%	30 (22, 38)	18 (14, 45), 28 (25.1, 30.9)	1-76	29 (5.0%), 1,003 (17.1%)
PPS 60-80%	23 (14, 32)	18 (5, 29), 43 (37.9, 48.1)	4-61	15 (2.6%), 453 (7.7%)

* Palliative Performance Scale.
** Confidence interval.
*** Second cell entries (where available), in parentheses, are for the Victoria Hospice patient cohort used to develop Prognostat.

ria Hospice cohort that was used to develop Prognostat. The table shows significant discrepancies in the distribution of percentages for age and cancer status. There is also a significant discrepancy in the distribution of percentages and median survival times for PPS.

Figure 1 / Kaplan-Meier Survival Curves by Initial Palliative Performance Scale (PPS) Score

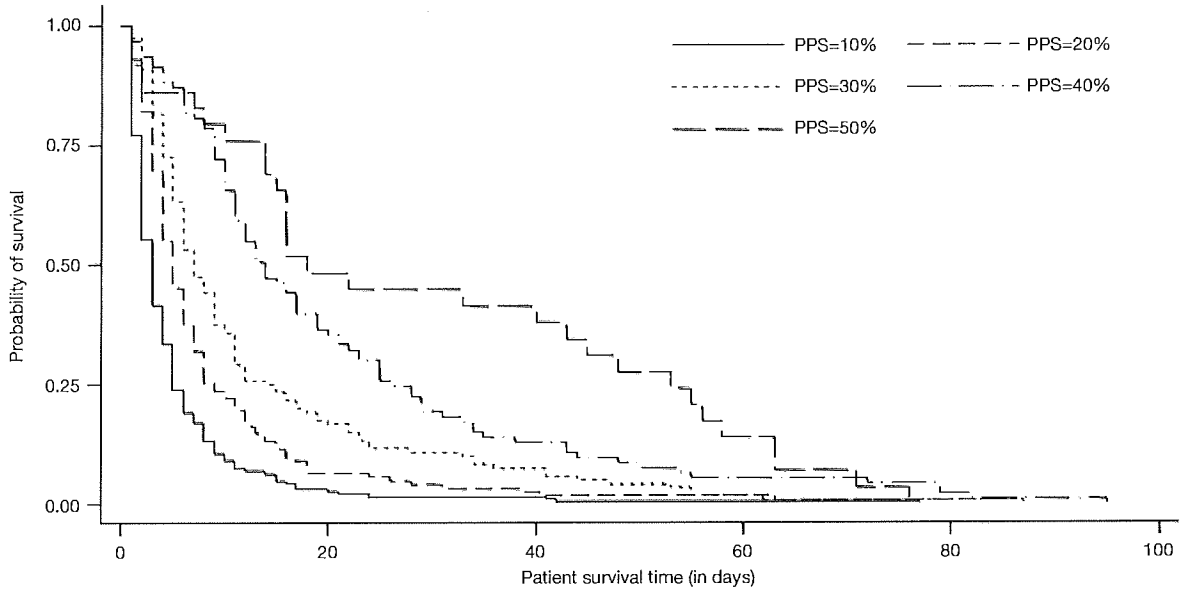


Table 2 / Accuracy, Discrimination, and Calibration Measures for Model Performance in Predicting Patient Survival Using Prognostat

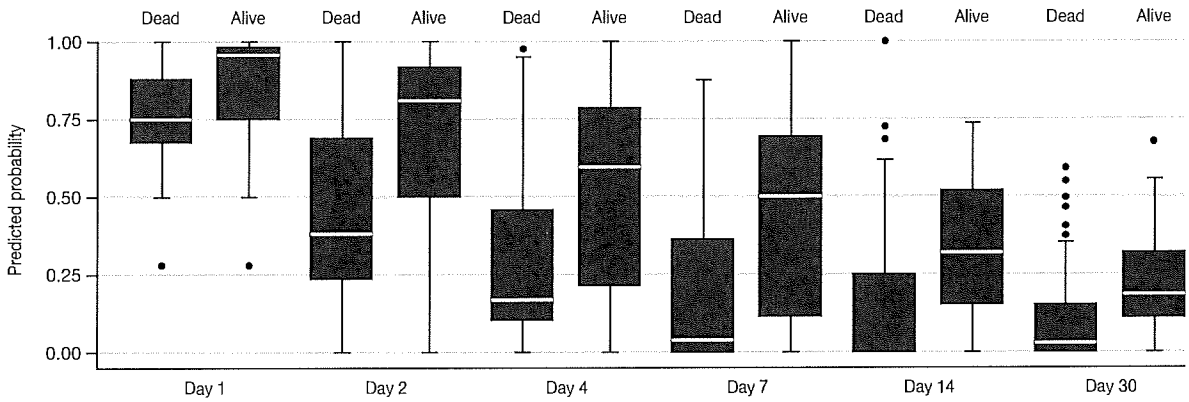
Day	% of patients survived	Score	PPS* adjusted
Day 1	89.4%	Brier Brier scaled AUC** (95% CI***) Slope HL**** p-value	0.099 18.30% 0.699 (0.628, 0.771) 0.119 0.0023
Day 2	77.6%	Brier Brier scaled AUC Slope HL p-value	0.187 18.40% 0.741 (0.689, 0.791) 0.234 < 0.001
Day 4	59.7%	Brier Brier scaled AUC (95% CI) Slope HL p-value	0.244 0.40% 0.724 (0.682, 0.767) 0.243 < 0.001
Day 7	42.4%	Brier Brier scaled AUC (95% CI) Slope HL p-value	0.211 2.20% 0.764 (0.724, 0.803) 0.263 < 0.001
Day 14	24%	Brier Brier scaled AUC (95% CI) Slope HL p-value	0.155 5.50% 0.756 (0.709, 0.804) 0.203 < 0.001
Day 30	10.8%	Brier Brier scaled AUC (95% CI) Slope HL p-value	0.088 39.40% 0.758 (0.696, 0.820) 0.136 < 0.001

* Palliative Performance Scale.
** Area under the receiver operating curve.
*** Confidence interval.
**** Hosmer-Lemeshow statistic.

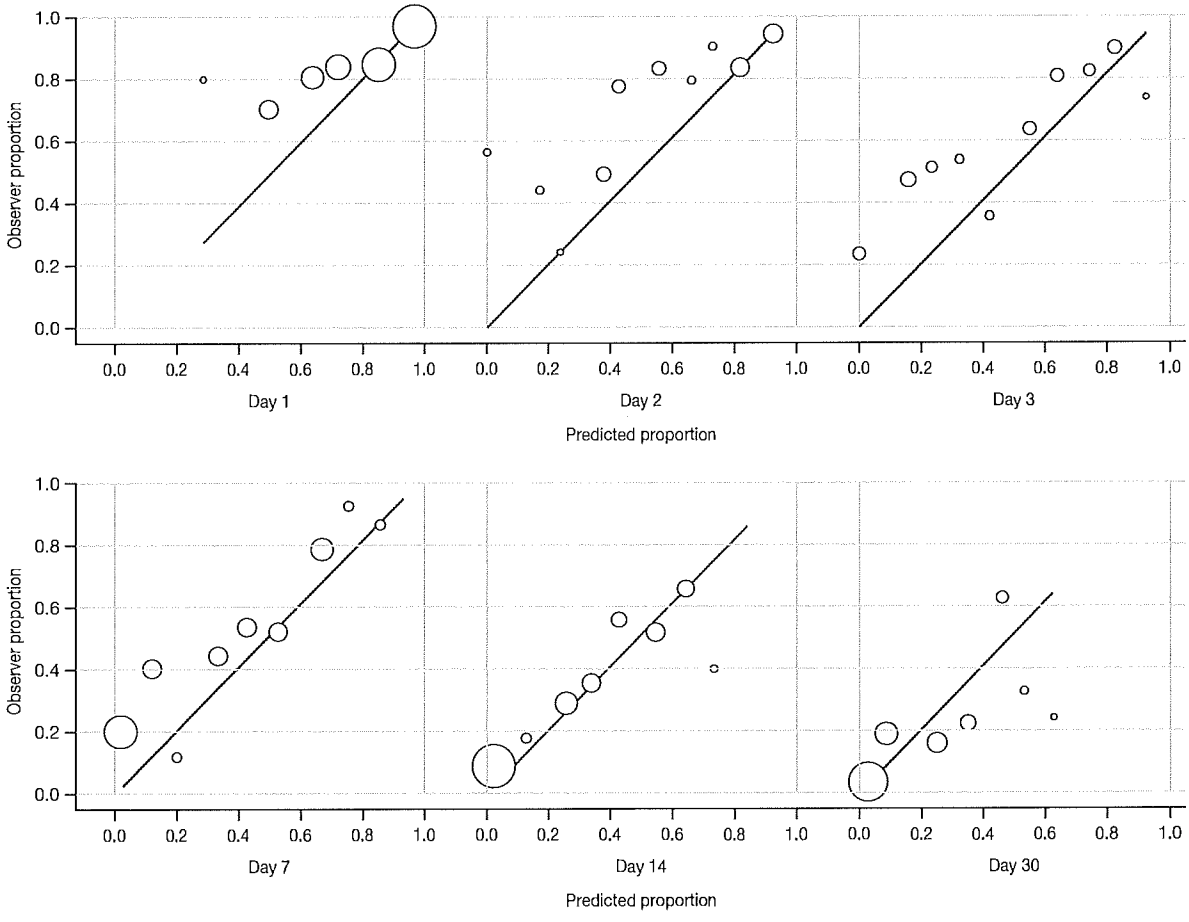
For our cohort, the Kaplan-Meier curves stratified by initial PPS level are shown in Figure 1. The curves show good separation, indicating that the different risk groups are well defined. We dropped 15 patients with PPS scores of 60 percent due to the crossing of the Kaplan-Meier estimate of PPS 50 percent. The log-rank test for equality of survival curves was highly significant at p=0.001 for PPS and cancer status, but not for age (p=0.303) and gender (p=0.944). Likewise, when adjacent categories of PPS were compared (PPS 10 percent versus 20 percent, 20 percent versus 30 percent, and so on), pairwise log-rank tests were all significant at p=0.05 level, except for PPS 40

Figure 2 / Predicted Versus Observed Patient Survival

2A: Box Plot Showing Predictions by Actual Outcome (Survival) for Hospice Patients



2B: Hosmer-Lemeshow Calibration Plot of Predicted Versus Actual Proportion of Survival
(The 45 degree line denotes the perfect agreement between predicted and observed survival.)



percent versus PPS 50 percent (p=0.394), due to initial crossing of two survival curves and the longer tail of the PPS 40 percent group. Patients who were 44 years old or younger did not have significantly lower hazard than those in the other age groups (p=0.862, 0.340, 0.466, 0.50, respectively), nor did male patients compared with female ones (p=0.806).

The measures of accuracy, discrimination, and calibration for days 1, 2, 4, 7, 14, and 30 are given in Table 2 and show poor performance of Prognostat overall. The discrimination slopes are relatively low and the Hosmer-Lemeshow goodness-of-fit test p-values are significant for all six days of measurement, indicating poor calibration. In the HL calibration plot of predicted versus observed proportion of those who survived (Figure 2B), circles are mostly unaligned with the 45-degree line. They show that in our cohort of patients, Prognostat consistently underestimates survival for days 1, 2, 4, 7, and 14, and overestimates it for day 30. The larger circles indicate that these points are based on more data. The absence of circles in any given decile indicates that there were no predictions in that interval. The overlapping box plots (Figure 2A) confirm poor discrimination.

DISCUSSION

This paper describes an external validation of the Web-based interactive prognostic tool Prognostat. We found that the tool performed poorly for our cohort of palliative patients. Since patient populations differ, it is not uncommon for the predictive performance of a model to deteriorate when the model is tested with patients other than those with whom it was developed. This has been recognized in the case of the PPS — due possibly to differences in patient cohort characteristics, location of care, and misunderstandings related to the use of the performance tool and the inter-reviewer discrepancy (18, 27). The differences between our cohort and the cohort used in the development of Prognostat are pronounced in terms age at treatment, cancer status, and PPS score.

However, we believe that instead of developing a new model, we should use knowledge from previous studies to update the existing prediction model by means of shrinkage and recalibration methods (28, 29). Updating methods can range from making adjustments to baseline survival to making adjustments to predictor weights using adjustment factors. This may entail re-estimating predictor weights and adding new predictors or removing existing predictors from the original model (10). Ideally, the updated model would also be externally validated. For Prognostat to be useful to hospice and palliative care researchers, it should report explicit risk scores to be combined with new patient information and provide guidance on how this should be done.

Prognostat is also restricted in the framework of the Cox proportional hazards model, especially due to the fact that it is impossible to directly model and report the baseline survival function. This is essential in calibrating survival estimates for a new population of patients. We have found that the Royston-Parmar family of survival functions (30) is more accurate and flexible than the Cox proportional hazards model (31), as it allows for parametric modelling of the baseline survival function and relaxing of the proportional hazards assumption.

LIMITATION

A limitation of our study is that it was confined to external validation of an existing model, which needs to be recalibrated and tested prospectively on a data set independent from our patient population. Without explicit information from Prognostat regarding patient risk scores and linear predictors, this is not feasible at this time.

ACKNOWLEDGEMENTS

This study was supported in part by a United States Army Medical Research and Materiel Command grant under the program Development of Evidence-Based Clinical Decision Support System to Aid Prognostication in Terminally Ill Patients (DOA W81 XWH-09-0175).

Received: September 5, 2012
Final version accepted: December 12, 2013

REFERENCES

1. Steinhauser KE, Christakis NA, Clipp EC, et al. Factors considered important at the end of life by patients, family, physicians, and other care providers. JAMA 2000; 284(19): 2476-2482.
2. Glare P, Virik K, Jones M, et al. A systematic review of physicians' survival predictions in terminally ill cancer patients. BMJ 2003; 327(7408): 195-198.
3. Tanneberger S, Malavasi I, Mariano P, et al. Planning palliative or terminal care: the dilemma of doctors' prognoses in terminally ill cancer patients. Ann Oncol 2002; 13(8): 1320-1322.
4. Viganò A, Dorgan M, Bruera E, et al. The relative accuracy of the clinical estimation of the duration of life for patients with end of life cancer. Cancer 1999; 86(1): 170-176.
5. Christakis NA, Lamont EB. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. BMJ 2000; 320(7233): 469-472.
6. Royston P, Moons S, Altman DG, et al. Prognosis and prognostic research: developing a prognostic model. BMJ 2009; 338: b604.
7. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000; 19(4): 453-473.
8. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010; 21(1): 128-138.
9. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. BMJ 2009; 338: b605.

10. Moons KG, Altman DG, Vergouwe Y, et al. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009; 338: b606.
11. Vickers AJ. Prediction models: revolutionary in principle, but do they do more good than harm? J Clin Oncol 2011; 29(22): 2951-2952.
12. University of Victoria Health Terminology Group. Prognostat. Accessed August 28, 2012. http://htg.his.uvic.ca/tools/PrognosticTools/PalliativePerformanceScale/Prognostat/index.php
13. Anderson F, Downing GM, Hill J, et al. Palliative performance scale (PPS): a new tool. J Palliat Care 1996; 12(1): 5-11.
14. Fainsinger RL, Demoissac D, Cole J, et al. Home versus hospice inpatient care: discharge characteristics of palliative care patients in an acute care hospital. J Palliat Care 2000; 16(1): 29-34.
15. Morita T, Tsunoda J, Inoue S, et al. Effects of high dose opioids and sedatives on survival in terminally ill cancer patients. J Pain Symptom Manage 2001; 21(4): 282-289.
16. Virik K, Glare P. Validation of the Palliative Performance Scale for inpatients admitted to a palliative care unit in Sydney, Australia. J Pain Symptom Manage 2002; 23(6): 455-457.
17. Morita T, Tsunoda J, Inoue S, et al. Validity of the Palliative Performance Scale from a survival perspective. J Pain Symptom Manage 1999; 18(1): 2-3.
18. Lau F, Bell H, Dean M, et al. Use of the Palliative Performance Scale in survival prediction for terminally ill patients in Western Newfoundland, Canada. J Palliat Care 2008; 24(4): 282-284.
19. Lau F, Maida V, Downing M, et al. Use of the Palliative Performance Scale (PPS) for end-of-life prognostication in a palliative medicine consultation service. J Pain Symptom Manage 2009; 37(6): 965-972.
20. Lau F, Downing M, Lesperance M, et al. Using the Palliative Performance Scale to provide meaningful survival estimates. J Pain Symptom Manage 2009; 38(1): 134-144.
21. Harrold J, Rickerson E, Carroll JT, et al. Is the Palliative Performance Scale a useful predictor of mortality in a heterogeneous hospice population? J Palliat Med 2005; 8(3): 503-509.
22. Downing M, Lau F, Lesperance M, et al. Meta-analysis of survival prediction with Palliative Performance Scale. J Palliat Care 2007; 23(4): 245-252.
23. Graf E, Schmoor C, Sauerbrei W, et al. Assessment and comparison of prognostic classification schemes for survival data. Stat Med 1999; 18(17-18): 2529-2545.
24. Hu B, Palta M, Shao J. Properties of R(2) statistics for logistic regression. Stat Med 2006; 25(8): 1383-1395.
25. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 2nd ed. New York: Wiley; 2000. (Wiley series in probability and statistics).
26. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143(1): 29-36.
27. Lau F, Downing GM, Lesperance M, et al. Use of Palliative Performance Scale in end-of-life prognostication. J Palliat Med 2006; 9(5): 1066-1075.
28. Janssen KJ, Moons KG, Kalkman CJ, et al. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol 2008; 61(1): 76-86.
29. Toll DB, Janssen KJ, Vergouwe Y, et al. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol 2008; 61(11): 1085-1094.
30. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med 2002; 21(15): 2175-2197.
31. Miladinovic B, Kumar A, Mhaskar R, et al. A flexible alternative to the Cox proportional hazards model for assessing the prognostic accuracy of hospice patient survival. PLoS One 2012; 7(10): e47804.
32. Royston P, Parmar MK, Altman DG. External validation and updating of a prognostic survival model. Department of Statistical Science, University College London; 2010 Mar. 17. Accessed June 1, 2013. http://www.ucl.ac.uk/statistics/research/pdfs/rr307.pdf
33. Royston P. Flexible parametric survival analysis using Stata: beyond the Cox model. College Station (TX): Stata Press; 2011.
34. Royston P. Flexible parametric alternatives to the Cox model, and more. Stata J 2001; 1(1): 1-28.
35. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. Stata J 2009; 9(2): 265-290.
36. Jackson C. Flexible parametric survival models. Cambridge (UK): author; 2013 May 15. http://cran.r-project.org/web/packages/flexsurv/index.html. Accessed January 31, 2012.

NOTE

[1] For a vector of covariates x and parameter vector β, the survival function S (t; x) for the Cox proportional hazards (CPH) model is commonly expressed as $S(t; x)=[S_0(t)]^{\exp(x\beta)}$ where S_0 (t) is the baseline survival function — that is, survival function when all the covariates x are equal to zero. In the CPH framework, the estimation of the (linear) prognostic index xβ does not require the formulation of the baseline cumulative survival function S_0 (t), which itself can be estimated conditional on the covariate estimates using the Breslow and Kalbfleisch-Prentice estimators. However, the full parametric estimation of S_0 (t) is not possible, which makes prediction of baseline survival from the primary to the secondary data set not viable.

An alternative to CPH is the Royston-Parmar family of survival models, which rely on the transformation g(.), such that $g(S(t; x))=g(S_0(t)) + x\beta$. The transformation g(.) can be either from the proportional hazard, proportional odds, Aranda-Ordaz, or probit families. The baseline survival function $S_0(t)$ is approximated and smoothed by a restricted cubic spline function with m interior knots. A desirable feature of these functions is that, unlike CPH, they can be reconstructed and used in postvalidation model calibrating if the scale used (hazard, probit, or odds), the knot positions, and the estimates of prognostic indices are reported. Calibration refers to estimating prognostic indices in the secondary data set using the parameter vector β estimated on the primary data set and applied to the vector of covariates x of the secondary data set. The interested reader is directed to a publication by Royston, Parmar, and Altman (32) for a detailed explanation. The methods can be implemented in Stata (33) statistical software using the *stpm* (34) and *stpm2* (35) commands, or in open source statistical software R as *flexsurv* package (36).

PLOS | ONE

A Flexible Alternative to the Cox Proportional Hazards Model for Assessing the Prognostic Accuracy of Hospice Patient Survival

Branko Miladinovic[1]*, Ambuj Kumar[1], Rahul Mhaskar[1], Sehwan Kim[2], Ronald Schonwetter[2], Benjamin Djulbegovic[1,3]

1 Center for Evidence Based Medicine and Health Outcomes Research, University of South Florida, Tampa, Florida, United States of America, 2 HPC Healthcare, Temple Terrace, Florida, United States of America, 3 H. Lee Moffitt Cancer Center & Research Institute, Tampa, Florida, United States of America

Abstract

Prognostic models are often used to estimate the length of patient survival. The Cox proportional hazards model has traditionally been applied to assess the accuracy of prognostic models. However, it may be suboptimal due to the inflexibility to model the baseline survival function and when the proportional hazards assumption is violated. The aim of this study was to use internal validation to compare the predictive power of a flexible Royston-Parmar family of survival functions with the Cox proportional hazards model. We applied the Palliative Performance Scale on a dataset of 590 hospice patients at the time of hospice admission. The retrospective data were obtained from the Lifepath Hospice and Palliative Care center in Hillsborough County, Florida, USA. The criteria used to evaluate and compare the models' predictive performance were the explained variation statistic R^2, scaled Brier score, and the discrimination slope. The explained variation statistic demonstrated that overall the Royston-Parmar family of survival functions provided a better fit ($R^2 = 0.298$; 95% CI: 0.236–0.358) than the Cox model ($R^2 = 0.156$; 95% CI: 0.111–0.203). The scaled Brier scores and discrimination slopes were consistently higher under the Royston-Parmar model. Researchers involved in prognosticating patient survival are encouraged to consider the Royston-Parmar model as an alternative to Cox.

Citation: Miladinovic B, Kumar A, Mhaskar R, Kim S, Schonwetter R, et al. (2012) A Flexible Alternative to the Cox Proportional Hazards Model for Assessing the Prognostic Accuracy of Hospice Patient Survival. PLoS ONE 7(10): e47804. doi:10.1371/journal.pone.0047804

Editor: Raya Khanin, Memorial Sloan Kettering Cancer Center, United States of America

Received May 11, 2012; **Accepted** September 21, 2012; **Published** October 17, 2012

Funding: This study was supported by the United States Army Medical Research and Material Command grant DOA W81 XWH-09-0175. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bmiladin@health.usf.edu

Introduction

Prognostic models are often used to estimate the length of patient survival and improvement in the accuracy of prognosis translates into superior quality of patient care. Precise prognosis of survival using modeling techniques requires rigorous methods for the development and testing of the accuracy of prognostic models. Developing a prognostic model entails having accurate patient data for prognosis, and selecting clinically relevant candidate predictors and measures of model performance, usually in the context of multivariable regression [1]. This process produces patient performance scores that allow for classification of patients into different risk groups [2,3,4].

In the hospice setting, accurate prognostication of survival affords patients and their families a vital opportunity to attend to matters such as planning, prioritizing, and preparing for death [5]. Predicting patient survival is a complex decision making process involving numerous subjective and numerical factors that have substantial variation which may lead to poor prediction of life expectancy. Many physicians practice optimism or avoidance, thus overestimating survival at times by a factor of five [6]. Implementing appropriate statistical methodologies translates into improved accuracy of prognosis and superior quality of care. Predictions based on appropriate statistical modeling have been shown to be superior to physicians' prognostication [4,7].

The Cox proportional hazards (CPH) model [8] is the most commonly-used survival prediction model [4,9]. In the hospice and palliative settings, demographic and clinical covariates are often included in CPH to predict patient survival [10,11]. The appeal of the model is its analytic simplicity and that the baseline survival function does not need to be defined *apriori*–it is absorbed when the likelihood function is maximized (note that "baseline" refers to zero values of the covariates, not to time equal to zero). It is possible to estimate the baseline survival function for the CPH model conditional on the estimated regression coefficients. However, this is highly rigid as the smoothing of the underlying function depends on the proportional hazards assumption, which may not be supported by the data and is often overlooked by the investigators [9]. Essentially, the CPH model was designed to measure the effects of covariates on the changing hazard function and not to model patient survival. A flexible family of functions which allows for parametrically modeling the baseline survival function is more appropriate, especially if the proportional hazards assumption is violated in the CPH [12]. The baseline survival has for the most part been ignored because it is left undefined in the CPH model.

In this manuscript we compare CPH with an alternative method of estimating survival in the form of the class of flexible Royston-Parmar (RP) parametric functions [12]. We use the Palliative Performance Scale (PPS) [13] from a cohort of hospice patients. Results from systematic reviews have shown that the patient PPS score is an accurate measure of patient survival in the palliative setting [7,11]. Furthermore, PPS and CPH model have been used to construct meaningful hospice patient survival estimates in the form of a life expectancy table and survival nomogram [14] and to validate prognosticating scales for hospice patient survival [15,16,17].

In addition to PPS, other risk factors such as age, cancer status and gender have been reported to be significant predictors of palliative patient survival in several studies [11,14]._ENREF_18 In our study we did not adjust for other risk factors because though they may be significant predictors of survival for the cohort of patients in our dataset, they may not be in other palliative settings. Our goal was to demonstrate that the RP family of parametric functions allowed for a direct and flexible modeling of the baseline survival and that it might be formulated so that the impact of the proportional hazard assumption is minimized. We determined if the overall performance and discriminatory ability of RP family of parametric functions is superior to CPH in the sample by using models that were derived and tested on the whole dataset (naïve validation) and using (internal) cross-validation. It is important to note that the RP parametric functions have not been applied to prognostic models in the hospice and palliative settings. It is also important to note that we did not perform external validation, which entailed using a different data set than the one used to create the model[3]. In the next section we briefly discuss PPS, introduce the statistical models and measures of model performance.

Methods

Study sample and palliative performance

The patient data were obtained from the Lifepath Hospice and Palliative Care Center licensed since 1983 to serve in Hillsborough County, Florida. Hospice care focuses on pain control and symptom management. To avoid selection bias, we retrospectively and sequentially extracted data for 590 patients who, as of January 2009 were deceased. This study was a retrospective review of the deceased patients' medical records. Only data pertaining to outcomes were collected; personal information was not collected and the data were de-identified prior to analysis. Since we did not collect any information that can identify deceased patients or their family members, under HIPPA rules and regulations (45 CFR 164.512) the requirement for consent does not apply. The study and consent procedures were approved by the University of South Florida Institutional Review Board prior to study initiation. Two research assistants extracted all data necessary to populate the model variables and two faculty members randomly checked 25% of the data for accuracy. The models were tested against observed survival duration.

The Palliative Performance Scale (PPS) was developed and reported by Anderson et al. [13] as a measure of palliative patients' functional status. The scale has 11 possible mutually exclusive levels, which are based on five domains: six levels of ambulation, six levels of activity and evidence of disease, five levels of self-care, five levels of food intake and four levels of consciousness. The scale ranges from PPS of 0% (deceased patient) to PPS of 100% (ambulatory and healthy patient). Numerous studies have studied its prognostic accuracy of survival in a variety of settings and found it provides meaningful estimates of patient survival [10,14,15,18,19,20,21,22,23]. PPS has been found to be both valid and reliable [24].

Model selection and validation

Validating a prognostic model is generally accepted to mean that given a patient population it works in a data set other than the one it is applied to[2,25]. In other words, the model needs to be tested using a different data set than the one used to create the model[3]. It is also generally accepted that the validation process should follow guidelines and that un-validated prognostic models should not be applied in clinical practice [3,4,26]. When validating a prognostic survival model in the regression framework, most attention has been on the value of the prognostic index based on covariates, while the role of the baseline survival function has been largely ignored.

The role of the baseline survival is significant as it quantifies the absolute patient survival probabilities over time. For a vector of covariates $\mathbf{x}$ and parameter vector $\boldsymbol{\beta}$, the survival function $S(t; \mathbf{x})$ at

Table 1. Patient characteristics.

Variable	Result
Total no. of patients	590 (100%)
Age at Treatment	
<45	37 (6.3%)
45–64	187 (31.7%)
65–74	110 (18.6%)
75–84	129 (21.9%)
85+	127 (21.5%)
Gender	
Male	293 (49.7%)
Female	295 (50%)
Unknown	2 (0.3%)
No. of patients with cancer/noncancer	
Noncancer	363 (61.5%)
Cancer	227 (38.5%)
Diagnosis category for cancer	
Brain	10 (1.7%)
Gastrointestinal	35 (5.9%)
Genital-female	12 (2%)
Genital-male	12 (2%)
Head and neck	8 (1.4%)
Hematopoietic	10 (1.7%)
Pancreas	24 (4.2%)
Respiratory	49 (8.3%)
Skin	2 (0.3%)
Urinary	4 (0.6%)
Other	61 (10.3%)
Diagnosis category for noncancer	
AIDS	12 (2%)
Cardiovascular	74 (12.5%)
Neurological	119 (20.2%)
Respiratory	37 (6.3%)
Other	121 (20.6%)

doi:10.1371/journal.pone.0047804.t001

Table 2. Survival time by age, gender, diagnosis and initial PPS.

Variable	Survival Times (in Days) Mean (95% CI)	Median (95% CI)	Range	No. of Patients (%)
Total no. of patients				
Overall	14 (12, 17)	6 (5, 6)	1–371	590
Age at Treatment				
<45	15 (8, 22)	8 (4,12)	1–95	37 (6.3%)
45–64	14 (11, 17)	7 (5, 9)	1–114	187 (31.7%)
65–74	14 (8, 20)	5 (4, 6)	1–271	110 (18.6%)
75–84	14 (8, 20)	6 (5, 7)	1–371	129 (21.9%)
85+	15 (9, 21)	5 (4, 6)	1–313	127 (21.5%)
Gender				
Male	14 (10, 18)	6 (5, 7)	1–371	293 (49.7%)
Female	15 (11, 19)	6 (5, 7)	1–271	295 (50%)
No. of patients with cancer				
Noncancer	12 (8, 16)	5 (4, 6)	1–371	363 (61.5%)
Cancer	17 (14, 20)	9 (7, 11)	1–113	227 (38.5%)
Diagnosis category for cancer				
Brain	27 (16, 39)	28 (14, 42)	3–55	10 (1.7%)
Gastrointestinal	21 (14, 29)	11 (5,17)	1–82	35 (5.9%)
Genital-female	15 (6, 24)	8 (1, 15)	2–55	12 (2%)
Genital-male	26 (7, 45)	13 (4, 22)	1–100	12 (2%)
Head and neck	10 (2, 18)	5 (1, 9)	1–36	8 (1.4%)
Hematopoietic	4 (2, 6)	3 (1, 5)	1–10	10 (1.7%)
Pancreas	18 (7, 29)	7 (3, 11)	1–113	24 (4.2%)
Respiratory	15 (10, 20)	10 (7, 13)	1–71	49 (8.3%)
Skin	11	11	11–11	2 (0.3%)
Urinary	25 (1, 58)	9 (1, 39)	4–76	4 (0.6%)
Other	17 (12, 22)	9 (5, 12)	1–103	61 (10.3%)
Diagnosis category for noncancer				
AIDS	18 (3, 33)	8 (1, 15)	1–85	12 (2%)
Cardiovascular	14 (5, 23)	5 (3, 7)	1–271	74 (12.5%)
Neurological	8 (5, 11)	5 (4,6)	1–77	119 (20.2%)
Respiratory	25 (1, 49)	3 (1, 5)	1–371	37 (6.3%)
Other	11 (1, 15)	5 (4, 6)	1–174	121 (20.6%)
Initial PPS Score				
PPS 10%	5 (3, 7)	3 (2, 4)	1–77	188 (32.6%)
PPS 20%	16 (8, 24)	5 (4, 6)	1–371	125 (21.7%)
PPS 30%	15 (11, 19)	7 (5, 9)	1–140	123 (21.4%)
PPS 40%	24 (18, 30)	14 (11, 17)	1–147	96 (16.7%)
PPS 50–80%	28 (21, 35)	18 (9, 27)	1–76	44 (7.6%)

doi:10.1371/journal.pone.0047804.t002

time t for the CPH model is commonly expressed as $S(t;x)=[S_0(t)]^{\exp(x\beta)}$, where S_0 (t) is the baseline survival function, i.e. survival function when all the covariates **x** are equal to zero. In the CPH framework, the estimation of the prognostic index $\mathbf{x}\boldsymbol{\beta}$ does not require the formulation of the baseline cumulative survival function S_0 (t), which itself can be estimated conditional on the covariate estimates. The two popular methods for estimating baseline survival S_0 (t) are the Breslow and Kalbfleisch-Prentice methods [27]. Both give similar results in practice, but can lead to "choppy" estimates of the baseline function and are dependent on the proportional hazards assumption.

When the goal of a survival analysis is to estimate hazard ratios (the effect of covariates on the changing hazard function), the baseline function is of no consequence. The CPH is appropriate as the baseline function gets absorbed when coefficient $\boldsymbol{\beta}$s are estimates by the method of partial log likelihood. However, when the goal is to prognosticate patient survival, there is a need for more flexibility in modeling the baseline survival.

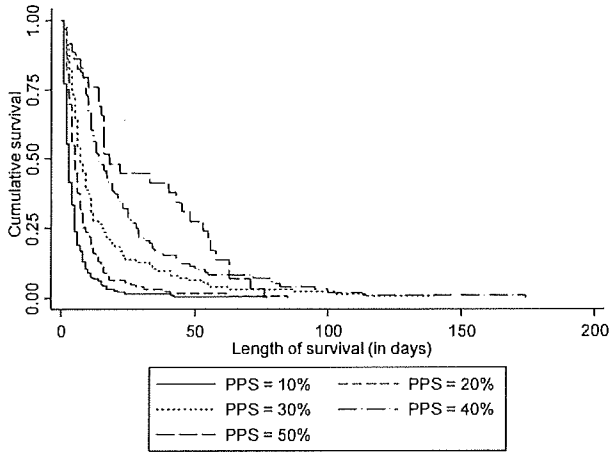


Figure 1. Kaplan-Meier survival curves by initial PPS.
doi:10.1371/journal.pone.0047804.g001

An alternative to the CPH is the RP family of models that resembles the generalized linear models and can be viewed as a parametric extension Cox proportional hazard models [12]. The models are framed to rely on the transformation g(.), such that $g(S(t;x))=g(S_0(t))+x\beta$. The transformation g(.) can be either from the proportional hazard, proportional odds, Aranda-Ordaz or probit families [12]. We did not consider the Aranda-Ordaz family in this study due to possible interpretational difficulties [12]. Under the proportional hazard link function, the hazard ratio estimates are nearly identical to those estimated under CPH. The attractive feature of the RP baseline survival function is that its shape is preserved, but the location of the baseline distribution function can vary, which allows for flexible model recalibration. Also, the estimate $g(s_0(t))$ is implemented on log-time scale. It is generally gently curved and smooth, making survival estimates more accurate.

In the RP framework, if the proportional hazard assumption is violated, the probit-link function $g(s)=-\Phi^{-1}(s)$ can be applied, where $\Phi^{-1}(.)$ is the inverse standard normal distribution function. The baseline survival function $s_0(t)$ is approximated and smoothed by a restricted cubic spline function with m interior knots. Splines are piecewise polynomials that ensure the overall curve is smooth (see Royston and Parmar [12] for details). Spline-based survival models such as RP have been empirically shown to be superior when the proportional hazard assumption is violated [28]. The optimal number of knots and the comparison among different RP models can be found using the minimum combination of Akaike Information Criterion (AIC), Bayes Information Criterion (BIC) and explained variation statistic R^2 [29,30]. The AIC is defined in the usual manner as - 2Log(likelihood) + 2(No. of model parameters), while BIC equals - 2Log(likelihood) + (No. of model parameters)*Log(n). In survival analysis n is interpreted as the number of events rather than the number of patients. The placement of knots in spline modeling is an issue. We have placed the knots at equally spaced centiles of the log-survival times, following published recommendations [31]. For example, for m = 1 the knot is at the 50^{th} centile, for m = 2 the knots are at the 33^{th} and 67^{th} centiles, etc.

We compared RP and CPH by performing internal validation (assessing validity in the population where the development data originated from) on the whole data set (naïve) and using split-sample cross-validation. We performed 10-fold cross-validation by splitting the data into development

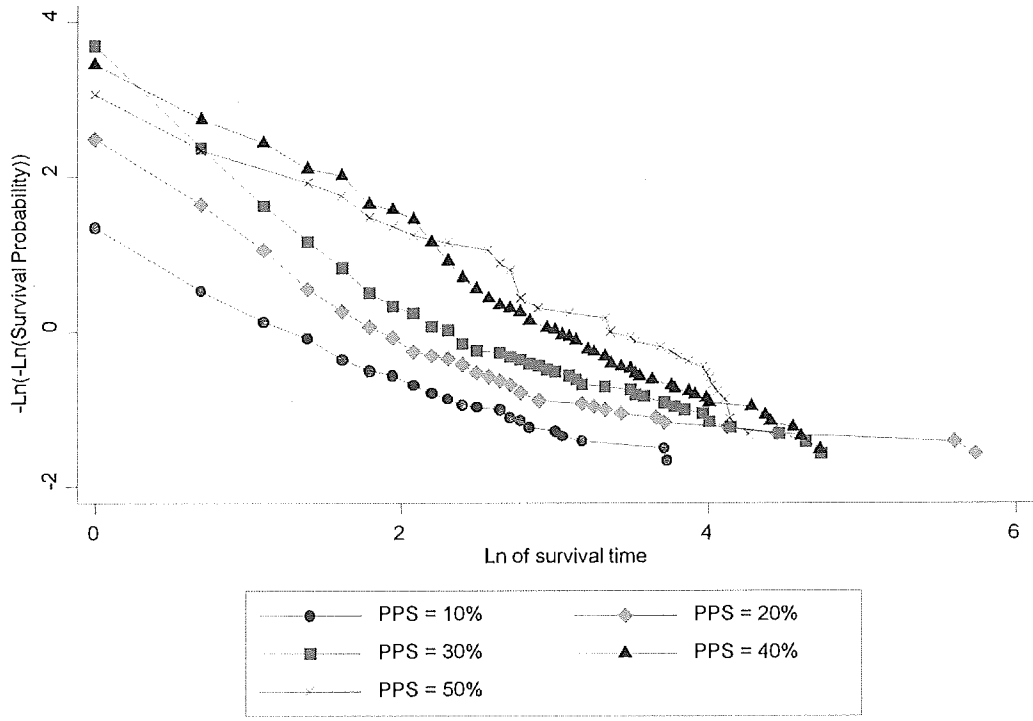


Figure 2. Test of the proportional hazards assumption under CPH for initial PPS.
doi:10.1371/journal.pone.0047804.g002

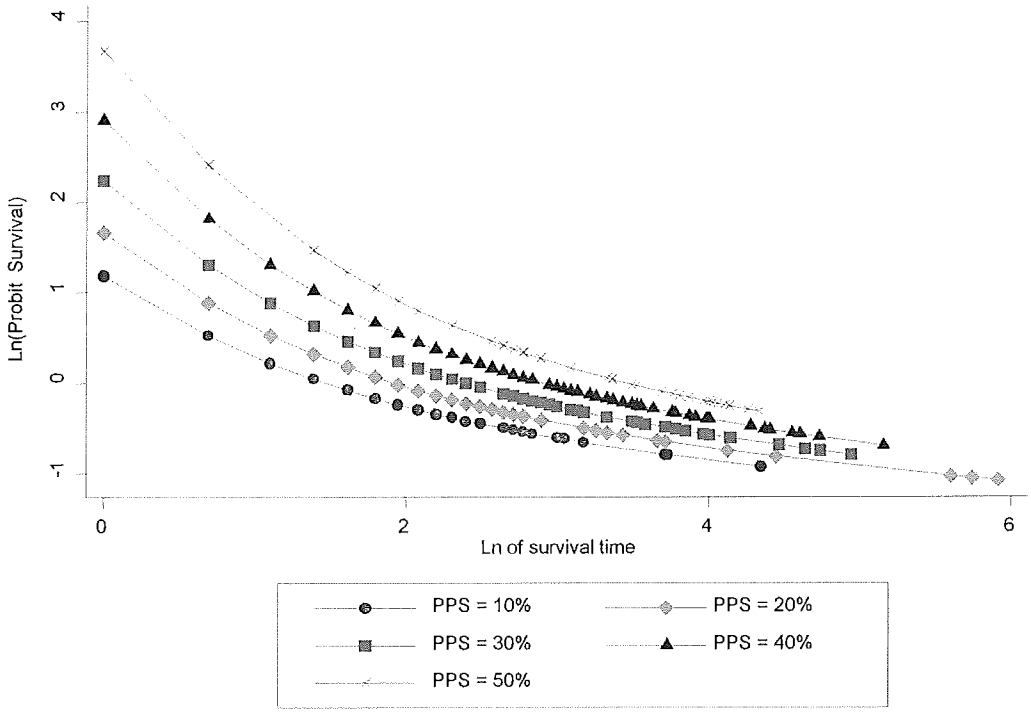


Figure 3. Test of the probit assumption under RP for initial PPS.
doi:10.1371/journal.pone.0047804.g003

and validation sets and repeating the process 20 times. The methods can be readily implemented in Stata [32,33] statistical software using the *stpm* [29] and *stpm2* [34] commands, or in open source statistical software R as *flexsurv* package [35].

Assessment of model performance

Model performance is the ability of the estimated risk score to predict survival and is assessed using the measures of explained variation, calibration, and discrimination. Calibration refers to how closely the predicted survival at a pre-specified time agrees with the observed survival. For cross-validation, we compared the average fitted probabilities of survival under RP and CPH for the first 15 days to observed probabilities estimated non-parametrically using Kaplan-Meier curves [36].

The Brier score is a quadratic scoring rule that calculates the differences between the actual outcomes and predicted probabilities[37]. Given the predicted probability of survival p_i at time t for patient i, and Y_i binary (0–1, dead-alive) variable, the Brier score is defined as $\sum(Y_i(1-p_i)^2+(1-Y_i)p_i^2)$. A Brier score of 0 indicates a perfect model, while 0.25 indicates a non-informative model (the value achieved when issuing a predicted probability of 50% to each patient). The Brier score may be scaled by its maximum $\text{Brier}_{\max}=(1 - \text{mean}(p_i))\ \text{mean}(p_i)$ to obtain $\text{Brier}_{\text{scaled}}=(1-\frac{\text{Brier}}{\text{Brier}_{\max}})100\%$. The scaled Brier scores range from 0% to 100% and have interpretation similar to the Pearson correlation coefficient[38].

Table 3. Criteria for the choice of scale in the RP model.

No. of knots m	PH	PO	Probit
	AIC, BIC, R^2	AIC, BIC, R^2	AIC, BIC, R^2
0	2033, 2042, 0.156	1887, 1896, 0.321	1872, 1881, 0.295
1	1889, 1902, 0.178	1883, 1896, 0.322	1858, 1871, 0.298
2	1871, 1888, 0.170	1870, 1887, 0.312	1857, 1874, 0.296
3	1870, 1892, 0.172	1870, 1892, 0.311	1858, 1880, 0.297
4	1865, 1892, 0.171	1865, 1891, 0.310	1855, 1881, 0.296
5	1866, 1896, 0.171	1865, 1896, 0.309	1856, 1886, 0.296

doi:10.1371/journal.pone.0047804.t003

For a particular risk score, discrimination is the ability to differentiate between the patients who died versus those who survived. The Kaplan-Meier plot of survival for patients in different risk groups can be used to test for separation, indicating that the different risk groups are well defined [39]. For a statistical model, the global measure of the model's discriminatory power is the explained variation statistic R^2, which measures the variation explained by the fitted model [40]. Higher values of R^2 indicate greater discrimination. In this study we implement R^2 for survival models, as described by Royston and Sauerbrei,[41].

The discrimination or Yates slope is a measure of how well the subjects with and without the outcome are separated. It is defined as the absolute difference in mean predictions of survival (mean[p_i]) between those who died and those who survived at time t[2]. The scaled Brier scores and discrimination slopes were calculated separately for the (naïve) model using the whole dataset and the model derived using cross-validation for t = 1, 2... 100 days. Higher scaled Brier scores and discrimination slopes represent better model performance.

All statistical calculation were performed using Stata version 11.2 [32,33].

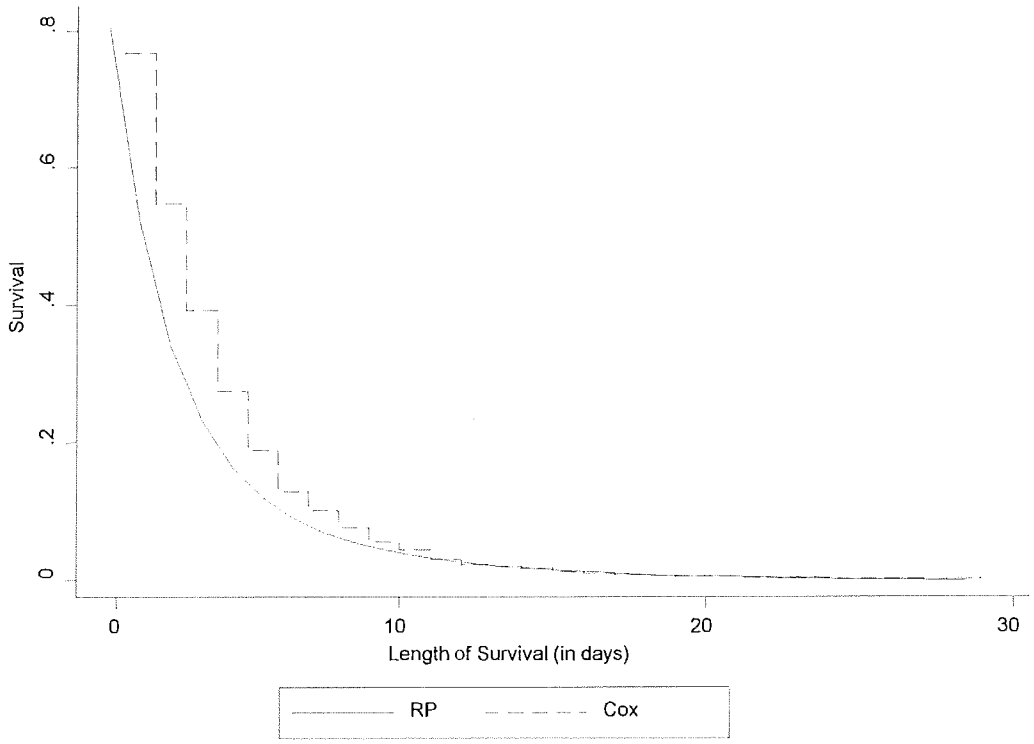


Figure 4. Baseline survival functions under CPH and RP models.
doi:10.1371/journal.pone.0047804.g004

Results

Description of the data source

The patient characteristics of the retrospective cohort are summarized in Table 1. The cohort consisted of 293 males (49.7%) and 295 females (50.0%), and 2 (0.3%) with unknown gender. The data were collected starting from patients' entry into hospice care until death for all 590 patients. The mean, median and range of survival times for the patients by PPS at admission, age, gender, cancer status, and diagnosis category are given in Table 2. The table shows that the median survival was fairly evenly distributed across age groups and gender, but unevenly across the cancer status and initial diagnosis category. All patients were assigned PPS at the time of

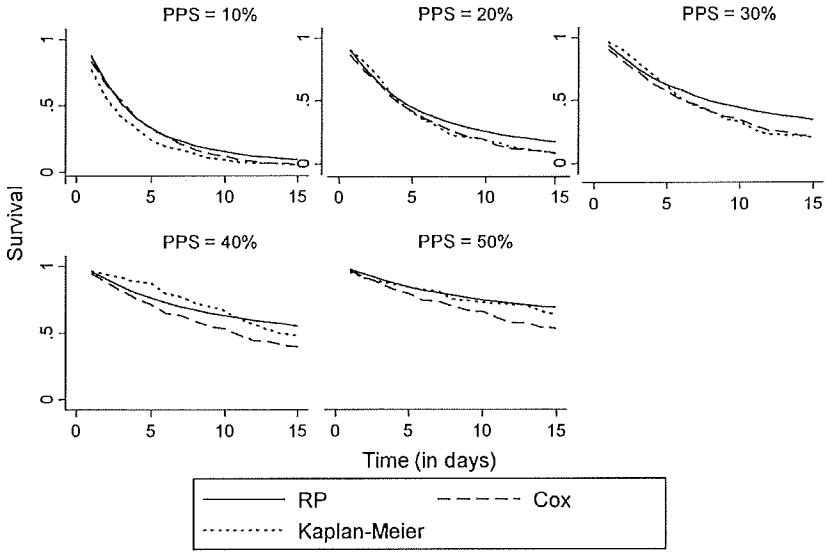


Figure 5. Predicted survival by PPS under RP and CPH compared with the Kaplan-Meier estimates in the validation data.
doi:10.1371/journal.pone.0047804.g005

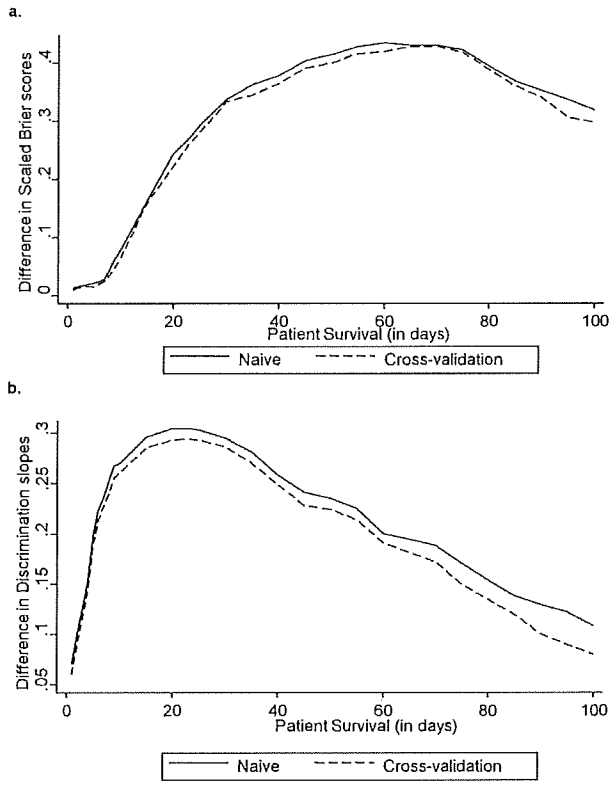


Figure 6. Difference between Brier scores for RP and CPH models (6a) and between discrimination slopes for RP and CPH models (6b) as a function of patient survival times in the naïve (whole) data set and cross-validated data set. Both are consistently higher for RP indicating better accuracy and discrimination.
doi:10.1371/journal.pone.0047804.g006

admission to hospice care. Since PPS score of 0% means that the patient is dead, the data were transformed so that the PPS score of 10% was set as the baseline. There were only 15 total observations for PPS = 60%, 70%, 80%, so they were combined with PPS = 50% to obtain meaningful survival estimates. Fourteen patients had missing values for PPS.

The time of admission was the starting point for survival time. The Kaplan-Meier curves stratified by initial PPS level are shown in Figure 1. The curves show good separation indicating that the different risk groups are well defined. The log-rank test for equality of survival curves was highly significant at $P = 0.001$. The global test based on Schoenfeld residuals showed that the proportional hazard assumption was violated for PPS (P-value <0.001), which can also be seen from the un-parallel natural log-plot of survival curves (Figure 2).

Table 3 lists AIC, BIC and R^2 values for 5 knots under the proportional hazard, proportional odds and probit RP families; the minimum combination in each is underlined. The number of optimal knots was found to be $m = 1$ under the probit model. The improvement in fit with the probit model can be seen from the parallel survival curves of log-probit against natural log time (Figure 3).

R^2 was higher in the RP model ($R^2 = 0.298$; 95% CI: 0.236–0.358) than the Cox model ($R^2 = 0.156$; 95% CI: 0.111–0.203), indicating that the RP model explained significantly more variation than CPH. To illustrate the differences for the baseline function, Figure 4 shows plots of the CPH and RP baseline survival functions. The CPH baseline survival is "choppy" to approximately day 12, while the RP is smooth. The two baseline functions converged at around day 12.

Cross-validation showed that the relation between the two predicted survival estimates is approximately linear, with RP model consistently estimating a higher probability, which is particularly evident for higher scores of PPS corresponding to longer survival times (Figure 5). Overall, the predicted probabilities under RP tended to be closer to the Kaplan-Meier estimates than CPH. The plot of the consistently positive differences between RP and CPH scaled Brier scores (Figure 6a) and discrimination slopes (Figure 6b) showed that the RP model discriminated better across patient survival times for both the full (naïve) and cross-validated models. This suggested that the higher value of R^2 under RP was not due to over-fitting.

Discussion

The results from our study show that RP family of models predicts survival more accurately than CPH through its flexible modeling of the baseline survival function. Using the RP flexible baseline function modeling would allow for more precise calibration in the prognostication phase than CPH. As Figure 5 illustrates, the predicted RP survival probabilities are consistently higher for higher values of PPS, and closer to the Kaplan-Meier estimates of survival. We suspect that both the robust modeling of baseline survival and overall model fit provide for better survival estimation.

There are limitations to our study, the primary one being the use of retrospective data. The RP family of parametric functions needs to be applied prospectively to assess accuracy of prognostic models through external validation. Furthermore, the dataset was limited to the hospice setting with no censored observations and with majority of patients having a very short follow-up time. For future studies, application of the proposed methodology should account for these limitations, and comparisons with parametric prognostic survival models should be explored.

The flexible models discussed in this paper could greatly improve the ability of researchers to accurately predict survival. An advantage of RP is that it can be used to validate published models for which the original individual patient data are unavailable. If the scale used (hazard, probit or odds), the knot positions, and the estimates of prognostic indices are known, then it would be possible to use RP. In the case of CPH this is not possible, since the baseline function would not be available.

Acknowledgments

The authors wish to thank Dr. Jane Carver for her help in preparing the manuscript.

Author Contributions

Conceived and designed the experiments: BM BD. Analyzed the data: BM. Contributed reagents/materials/analysis tools: RS SK. Wrote the paper: BM BD AK RM.

References

1. Royston P, Moons KG, Altman DG, Vergouwe Y (2009) Prognosis and prognostic research: Developing a prognostic model. BMJ 338: b604.

2. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, et al. (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21: 128–138.

3. Vickers AJ (2011) Prediction models: revolutionary in principle, but do they do more good than harm? J Clin Oncol 29: 2951–2952.
4. Moons KG, Altman DG, Vergouwe Y, Royston P (2009) Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 338: b606.
5. Steinhauser KE, Christakis NA, Clipp EC, McNeilly M, McIntyre L, et al. (2000) Factors considered important at the end of life by patients, family, physicians, and other care providers. JAMA 284: 2476–2482.
6. Christakis NA, Lamont EB (2000) Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. BMJ 320: 469–472.
7. Chow E, Harth T, Hruby G, Finkelstein J, Wu J, et al. (2001) How accurate are physicians' clinical predictions of survival and the available prognostic tools in estimating survival times in terminally ill cancer patients? A systematic review. Clin Oncol (R Coll Radiol) 13: 209–218.
8. Cox DR, Oakes D (1984) Analysis of survival data. London;New York: Chapman and Hall. viii, 201 p.p.
9. Mallett S, Royston P, Waters R, Dutton S, Altman DG (2010) Reporting performance of prognostic models in cancer: a review. BMC Med 8: 21.
10. Downing M, Lau F, Lesperance M, Karlson N, Shaw J, et al. (2007) Meta-analysis of survival prediction with Palliative Performance Scale. J Palliat Care 23: 245–252; discussion 252–244.
11. Lau F, Cloutier-Fisher D, Kuziemsky C, Black F, Downing M, et al. (2007) A systematic review of prognostic tools for estimating survival time in palliative care. J Palliat Care 23: 93–112.
12. Royston P, Parmar MK (2002) Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med 21: 2175–2197.
13. Anderson F, Downing GM, Hill J, Casorso L, Lerch N (1996) Palliative performance scale (PPS): a new tool. J Palliat Care 12: 5–11.
14. Lau F, Downing M, Lesperance M, Karlson N, Kuziemsky C, et al. (2009) Using the Palliative Performance Scale to provide meaningful survival estimates. J Pain Symptom Manage 38: 134–144.
15. Harrold J, Rickerson E, Carroll JT, McGrath J, Morales K, et al. (2005) Is the palliative performance scale a useful predictor of mortality in a heterogeneous hospice population? J Palliat Med 8: 503–509.
16. Olajide O, Hanson L, Usher BM, Qaqish BF, Schwartz R, et al. (2007) Validation of the palliative performance scale in the acute tertiary care hospital setting. J Palliat Med 10: 111–117.
17. Head B, Ritchie CS, Smoot TM (2005) Prognostication in hospice care: can the palliative performance scale help? J Palliat Med 8: 492–502.
18. Fainsinger RL, Demoissac D, Cole J, Mead-Wood K, Lee E (2000) Home versus hospice inpatient care: discharge characteristics of palliative care patients in an acute care hospital. J Palliat Care 16: 29–34.
19. Morita T, Tsunoda J, Inoue S, Chihara S (2001) Effects of high dose opioids and sedatives on survival in terminally ill cancer patients. J Pain Symptom Manage 21: 282–289.
20. Virik K, Glare P (2002) Validation of the palliative performance scale for inpatients admitted to a palliative care unit in Sydney, Australia. J Pain Symptom Manage 23: 455–457.
21. Morita T, Tsunoda J, Inoue S, Chihara S (1999) Validity of the palliative performance scale from a survival perspective. J Pain Symptom Manage 18: 2–3.
22. Lau F, Bell H, Dean M, Downing M, Lesperance M (2008) Use of the Palliative Performance Scale in survival prediction for terminally ill patients in Western Newfoundland, Canada. J Palliat Care 24: 282–284.
23. Lau F, Maida V, Downing M, Lesperance M, Karlson N, et al. (2009) Use of the Palliative Performance Scale (PPS) for end-of-life prognostication in a palliative medicine consultation service. J Pain Symptom Manage 37: 965–972.
24. Ho F, Lau F, Downing MG, Lesperance M (2008) A reliability and validity study of the Palliative Performance Scale. BMC Palliat Care 7: 10.
25. Altman DG, Royston P (2000) What do we mean by validating a prognostic model? Stat Med 19: 453–473.
26. Altman DG, Vergouwe Y, Royston P, Moons KG (2009) Prognosis and prognostic research: validating a prognostic model. BMJ 338: b605.
27. Kalbfleisch JD, Prentice RL (2002)The statistical analysis of failure time data. Hoboken, N.J.: J. Wiley.. xiii, 439 p.p.
28. Binquet C, Abrahamowicz M, Mahboubi A, Jooste V, Faivre J, et al. (2008) Empirical study of the dependence of the results of multivariable flexible survival analyses on model selection strategy. Stat Med 27: 6470–6488.
29. Royston P (2001) Flexible parametric alternatives to the Cox model, and more. Stata Journal 1: 1–28.
30. Schwarz G (1978) Estimating Dimension of a Model. Annals of Statistics 6: 461–464.
31. Durrleman S, Simon R (1989) Flexible regression models with cubic splines. Stat Med 8: 551–561.
32. Stata Version 11 [computer program]. 9 ed. College Station, TX: Stata Corporation; 2010.
33. Royston P (2011) Flexible parametric survival analysis using stata : beyond the Cox model. College Station, TX: Stata Press.
34. Lambert PC, Royston P (2009) Further development of flexible parametric models for survival analysis. The Stata Journal 9: 265–290.
35. Jackson C (2012) Flexible parametric survival models.
36. Cook NR, Buring JE, Ridker PM (2006) The effect of including C-reactive protein in cardiovascular risk prediction models for women. Ann Intern Med 145: 21–29.
37. Graf E, Schmoor C, Sauerbrei W, Schumacher M (1999) Assessment and comparison of prognostic classification schemes for survival data. Stat Med 18: 2529–2545.
38. Hu B, Palta M, Shao J (2006) Properties of R(2) statistics for logistic regression. Stat Med 25: 1383–1395.
39. Altman DG (2009) Prognostic models: a methodological framework and review of models for breast cancer. Cancer Invest 27: 235–243.
40. Royston P (2006) Explained variation for survival models. The Stata Journal 6: 1–14.
41. Royston P, Sauerbrei W (2004) A new measure of prognostic separation in survival data. Stat Med 23: 723–748.

Running head: LUNG CANCER PROGNOSIS

Running head: LUNG CANCER PROGNOSIS

Natural History of Patients With Lung Cancer Without Treatment: A systematic Review

Hesborn Wao, PhD[1]
Rahul Mhaskar, PhD[1]
Benjamin Djulbegovic, MD, PhD[1,2]
Ambuj Kumar, MD, MPH[1,2]

Author Affiliations:

1. Center for Evidence Based Medicine and Health Outcomes Research, University of South Florida, Tampa, Florida, USA.

2. Moffitt Cancer Center & Research Institute, Tampa, Florida, USA.

Corresponding Author

Hesborn Wao, Ph.D.
Assistant Professor
Center for Evidence-Based Medicine and Health Outcomes Research
USF Health Morsani College of Medicine, University of South Florida
12901 Bruce B Downs Blvd., MDC27 Tampa, Florida 33612
Email: hwao1@health.usf.edu | Tel: (813)974-9248| Fax: (813)905-8909
http://health.usf.edu/research/ebm/index.htm

Rahul Mhaskar[1]
Email: rmhaskar@health.usf.edu

Benjamin Djulbegovic [1,2]
Email: bdjulbeg@health.usf.edu

Ambuj Kumar[1,2]
Email: akumar1@health.usf.edu

Survival of Patients With Non-Small Cell Lung Cancer Without Treatment: A Systematic Review and Meta-Analysis

Abstract

Background

Lung cancer is considered a terminal illness with a 5-year survival rate of about 16%. Informed decision related to management of a disease requires accurate prognosis of the disease with or without treatment. Despite the significance of disease prognosis in clinical decision-making, systematic assessment of prognosis in patients with lung cancer without treatment has not been performed. We conducted a systematic review and meta-analysis of the natural history of patients with confirmed diagnosis of lung cancer without active treatment, to provide evidence-based recommendations for practitioners on management decisions related to the disease. Specifically, we estimated overall survival when no anticancer therapy is provided.

Methods

Relevant studies were identified by search of electronic databases and abstract proceedings, review of bibliographies of included articles, and contacting experts in the field. All prospective or retrospective studies assessing prognosis of lung cancer patients without treatment were eligible for inclusion. Data on mortality was extracted from all included studies. Pooled proportion of mortality was calculated as a back-transform of the weighted mean of the transformed proportions, using the random-effects model. To perform meta-analysis of median survival, published methods were used to pool the estimates as mean and standard error under the random effects model. Methodological quality of the studies was examined.

Results

Seven cohort studies (4,418 patients) and 15 randomized controlled trials (1,031 patients) were included in the meta-analysis. All studies assessed mortality without treatment in patients with non-small cell lung cancer (NSCLC). The pooled proportion of mortality without treatment in cohort studies was 0.97 (95% CI: 0.96 to 0.99) and 0.96 in randomized controlled trials (95% CI: 0.94 to 0.98) over median study periods of 8 and 3 years, respectively. When data from cohort and randomized controlled trials were combined, the pooled proportion of mortality was 0.97 (95% CI 0.96 to 0.98). Test of interaction showed a statistically non-significant difference between subgroups of cohort and randomized controlled trials. The pooled mean survival for patients without anticancer treatment in cohort studies was 11.94 months (95%CI: 10.07 to 13.8) and 5.03 months (95%CI: 4.17 to 5.89) in RCTs. For the combined data (cohort studies and RCTs), the pooled mean survival was 7.15 months (95%CI: 5.87 to 8.42), with a statistically significant difference between the two designs. Overall, the studies were of moderate methodological quality.

Conclusion

Systematic evaluation of evidence on prognosis of NSCLC without treatment shows that mortality is very high. Untreated lung cancer patients live on average for 7.15 months. Although limited by study design, these findings provide the basis for future trials to determine optimal expected improvement in mortality with innovative treatments.

Keywords: Best supportive care, Natural history, Meta-analysis, Palliative care, Placebo

Background

Cancer is a major public health concern globally. It is the most frequent cause of death in economically developed countries.[1] Among all cancers, lung cancer is the leading cause of cancer deaths worldwide. [2] In the United States, approximately 221,130 new cases of lung cancer (14% of all cancer diagnoses) are expected in 2011 out of which 156,940 deaths (27% of cancer deaths) are estimated due to lung cancer.[3] Given the incurative nature of lung cancer, it is considered a terminal illness with a 5-year survival rate of approximately 16% .[3]

Patients diagnosed with terminal illness such as lung cancer confront several decisions related to management of the disease. Opting for treatment (e.g. chemotherapy, radiotherapy, or surgery) instead of palliation or vice versa is one such critical decision. Depending on the stage of the disease, potential benefits of anticancer therapy intended to palliate specific tumor-related symptoms may be at the expense of treatment-related harms and the inconvenience associated with undergoing treatment. Other times, palliative care (e.g. pain medications or low dose radiotherapy)[4] rather than anticancer therapy may be preferable. Informed decision related to management of a terminal disease thus requires accurate prognosis of the disease with or without treatment.

Briefly, prognosis refers to the likelihood of an individual developing a particular health outcome over a given period of time, based on the individual's clinical and non-clinical profile.[5] Accurate assessment of prognosis is key to informed decision making. For example, if a patient is diagnosed with a terminal illness such as lung cancer, a prognostic question of critical concern to the patient, family, and the physician is how long the patient is expected to live. Other

important outcomes may include disease progression, health-related quality of life, and treatment-related harms. Reliable prognostication of life expectancy can prevent subjecting patients to costly and unnecessary treatment for an unduly long period before transitioning to hospice care.[6] This in turn can help patients and their families prepare for the impending events and plan for the patient's remaining lifespan.[7] Accurate prognostic information can also help physicians decide on choice of curative versus palliative treatments. For instance, if evidence shows no effect of curative treatment on disease progression, significant treatment-related harms can be avoided in favor of palliative treatments.[7] It can help investigators avoid *optimism bias*, the "unwarranted belief in the efficacy of new therapies" [8] or making "overly optimistic assumptions regarding treatment benefits when designing RCTs."[9] Accurate disease prognosis thus underpins all management decisions related to the disease including choice of treatment, planning of supportive care, as well as allocation of resources.

Despite the significance of disease prognosis in clinical decision-making, systematic assessment of prognosis in patients with lung cancer without treatment has not been performed. We are aware of only one narrative review on the subject.[10] Accordingly, this systematic review was undertaken to assess the survival of patients with confirmed diagnosis of lung cancer without active treatment. Specifically, our aim was to estimate overall survival in lung cancer when no anticancer therapy is provided.

Methods

This systematic review was conducted as per the methods elaborated in a protocol that was developed *a priori.* The results are reported according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement.[11] An ideal study design to assess

natural history of a terminal disease such as lung cancer is a cohort study. Specifically, an inception cohort whereby a well-defined group of patients at the same disease stage is assembled at first diagnosis and followed for a defined period of time.[12-14] However, given the availability of treatments for lung cancer in recent years, it would be unethical and logistically challenging to conduct such a study. An alternative approach is to assess prognosis from retrospective lung cancer registries, case series or from the control arm of individual RCTs that compare active treatment with either no treatment or placebo or best supportive care.[5, 15].

Study eligibility

In this review, any retrospective or prospective cohort study assessing prognosis in lung cancer without treatment and any RCT assessing the role of treatment versus no treatment, were eligible for inclusion. A study was eligible for inclusion irrespective of language or publication type.

Search strategy

We conducted a systematic search of MEDLINE and Cochrane library electronic databases, proceedings of major scientific meetings, and bibliographies of eligible studies to identify all relevant studies. To retrieve lung cancer prognosis studies in PubMed, we employed search strategies suggested by Wilczynski[16] that optimizes search sensitivity and specificity. Search details used included: ("lung neoplasms"[MeSH Terms] AND "prognosis"[All Fields] AND "cohort"[All Fields] AND ("mortality"[Subheading] OR "natural course"[All Fields] OR "mortality"[All Fields] OR "survival"[All Fields] OR "survival"[MeSH Terms]). To retrieve RCTs in PubMed, we employed strategies suggested by Haynes[17] with the following search details: ("lung neoplasms"[MeSH Terms] AND ("randomized controlled trial"[Publication Type]) AND ("palliative care"[All Fields] OR "hospice care"[All Fields] OR "supportive care"[All Fields] OR "best supportive care"[All Fields] OR "placebo"[All Fields] OR

"symptomatic treatment"[All Fields] OR "no chemotherapy"[All Fields] OR "no treatment"[All Fields]). In the Cochrane library, we utilized a free text search using the term "Lung cancer" to identify RCTs focusing on lung cancer. We manually searched abstracts of the American Society of Clinical Oncology and American Society of Hematology meetings and utilized the snowballing procedure to identify other relevant studies. Studies published until June 2011 were included. No restrictions were made regarding the language of the publication.

Inclusion and exclusion criteria

A prospective or retrospective cohort study assessing overall survival as an outcome in lung cancer patients without treatment was eligible for inclusion. A RCT was included if it enrolled patients with confirmed diagnosis of lung cancer, compared treatment versus no treatment (e.g. supportive care, best supportive care, palliative care, placebo etc.), and assessed overall survival as an outcome. A study in which patients had anticancer treatment prior to enrollment and subgroup analyses were excluded. Additionally, RCTs comparing two active treatments were excluded. Two reviewers read the titles and abstracts of identified citations to identify potentially eligible studies. Full text of potentially relevant reports were retrieved and examined for eligibility. Disagreements about study inclusion or exclusion were resolved via discussion until a consensus was reached.

Data extraction

Data extraction was performed using a standardized data extraction form. Two reviewers independently extracted the following information from each included study: number of patients enrolled, number of deaths, median survival, funding source (industry versus public etc.), type of centers involved (single versus multicenter etc.), patient demographics, patients baseline clinical characteristics, and type of control arm (for RCTs only). For cohort studies, we extracted data on

the number of deaths and total number of patients diagnosed with lung cancer. For RCTs, we extracted data on the number of deaths (all-cause mortality) and number of participants randomized to the control arm.

Assessment of methodological quality

To evaluate the methodological quality of included studies, a modified checklist of predefined criteria was developed on four methodological domains pertinent to minimization of bias. This modified checklist uses applicable elements from existing tools (Quality in Prognosis Studies tool,[18] Evidence-Based Medicine Group criteria for prognostic studies,[19] Newcastle-Ottawa Quality Assessment Scale,[31] and Cochrane Collaboration risk of bias criteria[20]) and related studies (Hudak et al[21] and Altman[22]). The four domains included *participation bias* (extent to which study sample represents the population of interest on key characteristics), *attrition bias* (extent to which loss to followup of the sample was not associated with key characteristics), *outcome measurement* (extent to which outcome of interest is adequately measured in study participants), *data analysis* and *reporting* (extent to which statistical analysis and data reporting are appropriate for the study design). The modified checklist contains 11 items for cohort studies and 14 items for RCTs. For each item, a study either fulfilled a certain criterion (scored "Yes") or failed to fulfill the criterion (scored "No"). To assess methodological quality of studies included, we focused on proportion of studies that fulfilled each quality criterion (Table 2).

Statistical analysis

Data synthesis was conducted according to the study design separately as well as combined in the final stage (i.e., retrospective cohort and RCT). For the purpose of meta-analysis, we used methods by Stuarts et al[23] to transform the proportions into a quantity according to the Freeman-Tukey variant of the arcsine square root transformed proportion. The pooled proportion

was calculated as a back-transform of the weighted mean of the transformed proportions, using the random-effects model. To perform meta-analysis of median survival, we used published methods[24] to pool the estimates as mean survival and standard error under the random effects model. That is, using median survival and range reported in Kaplan-Meier curve, we converted these estimates into mean survival and standard error. Heterogeneity of treatment effects between trials was assessed using the I^2 statistic[20] with the following thresholds for I^2 statistic values: low (25% to 49%), moderate (50% to 74%), and high ($\geq$ 75%).[25] We explored the potential causes of heterogeneity by assessing the differences between subgroups using the test of interaction. We assessed robustness of the results by conducting sensitivity analysis with respect to methodological quality criteria of reporting, study location, and funding source. RevMan Version 5.1[26] was used to perform the analyses.

Results

Literature search

A flow diagram depicting the literature search process based on PRISMA [11] is shown in Figure 1. Initial search identified 1,562 potentially relevant citations excluding 71 duplicates. After initial screening of titles and abstracts, 1,489 records were not relevant for reasons depicted in Figure 1 and were excluded. Further assessment of full texts of remaining 73 studies led to exclusion of 51 studies. Altogether, 22 studies met the pre-defined inclusion criteria: 7 were retrospective cohort studies [27-33] and 15 were RCTs.[34-48]

Insert **Figure 1** about here

Study characteristics

We did not find any inception cohort study or a prospective cohort study assessing prognosis of patients with lung cancer without treatment. The seven retrospective cohort studies included 4,418 patients and the 15 RCTs enrolled 1,031 patients. Altogether, the 22 studies included 5,449 patients. All studies assessed prognosis in patients with NSCLC and were published between 1973 and 2009 (Table 1).

Cohort Studies: The median sample size in the cohort studies was 131 patients (range: 39 to 2,344 patients) with a median study period of 8 years (range: 5 to 13 years). Fifty-seven percent (4/7) and 29% (2/7) of the studies reported number of patients with stage I and stage II NSCLC, respectively. Forty-three percent (3/7) of the studies reported patients' cancer histology. Seventy-one percent (6/7) of the studies reported patient's gender. Forty-three percent (3/7) of the studies reported median age. Forty-three percent (3/7) of the studies were conducted at single institutions, 43% (3/7) were at multicenter national studies, and 14% (1/7) of the studies had unspecified study location. Twenty-nine percent (2/7) of the studies were publicly funded, 14% (1/7) were funded by both public and industry, and 57% (4/7) had not specified funding sources.

RCTs: The median number of patients enrolled in the RCTs was 61 patients (range: 17 to 176 patients) with a median study period of 3 years (range: 1 to 7 years). Median follow-up was reported in 33% (5/15 of RCTs) and ranged between 2.7 and 43 months. Seventy-three percent (11/15) of the studies reported number of patients with stage III/IV NSCLC. Seventy-three percent (13/15) of the studies reported patients' cancer histology. Eighty-seven percent (13/15) of the RCTs reported patient's gender and median age. Twenty percent (3/15) of the RCTs were conducted at single institutions, 27% (4/15) were at multicenter national studies, 20% (3/15)

were at multicenter international, and 33% (5/15) had unspecified study location. Seven percent (1/15) of the RCTs were funded by public, 33% (5/15) were funded by industry, 7% (1/15) were funded both public and industry, and 53% (8/15) had unspecified funding sources.

Types of control in RCTs: Three studies described *best supportive care* as comprising "symptomatic or palliative treatment excluding chemotherapy,"[49] "palliative radiotherapy, antibiotics, and corticosteroids,"[35] "palliative radiotherapy, opioid analgesics, and psychosocial support,"[42] or "radiation therapy, pain medication, nutritional and psychological support, thoracocentesis and/or tube thorascopy."[48] Three studies described *supportive care* as comprising "analgesics, an antitussive, relief of increased intracranial pressure, palliative radiotherapy, treatment of infections and pleural effusions,"[35] "symptomatic irradiation to involved fields,"[36] or "palliative radiation, analgesics, and psychosocial/nutritional support."[40] *Palliative care* consisted of "radiotherapy, antibiotics, coughs suppressants, and analgesics"[38] *Symptomatic treatment* included "glucocorticosteroids and anabolic steroids." [43] No descriptions were provided for *placebo* and "*no treatment.*"

Insert **Table 1** about here

Methodological quality

Cohort: All seven cohort studies fulfilled 64% (7/11) of the quality criteria (Table 2). That is, adequate description of population of interest for key characteristics, adequate description of study setting/geographic location, adequate participation in the study by all eligible patients, reporting of patients with missing data, a priori and objective definition of outcomes, and

presentation of frequencies of most important data (e.g., outcome) were reported in all studies. However, baseline sample was adequately described for key characteristics in 57% (4/7) of the studies, inclusion and exclusion criteria were adequately described in 71% (5/7) of the studies, follow-up was sufficiently long for outcome to occur in 86% (6/7) of the studies, and alpha error and/or beta error were specified *a priori* in 29% (2/7) of the studies.

RCTs: All 15 RCTs fulfilled 36% (5/14) of the quality criteria (Table 2). That is, adequate description of population of interest for key characteristics, adequate description of withdrawal (incomplete outcome data), a priori and objective definition of outcomes, and frequencies of most important data were reported in all RCTs. However, study setting and geographic location were adequately described in 47% (7/15) of the RCTs, baseline sample was adequately described for key characteristics in 93% (14/15) of the RCTs, inclusion and exclusion criteria were adequately described in 93% (14/15) of the RCTs, patients were balanced in all aspects except the intervention in 93% (14/15) of the RCTs, follow-up was sufficiently long for outcome to occur in 53% (8/15) of the RCTs, proportion of sample completing the study was adequate in 60% (9/15) of the RCTs, characteristics of dropouts versus completers was provided in 13% (2/15) of the RCTs, alpha error and/or beta error was specified a priori in 47% (7/15) of the RCTs, and data analysis was based on intention to treat analysis principle in 53% (9/15) of the RCTs.

Insert **Table 2** about here

Mortality

Cohort: Data on mortality was extractable from all seven cohort studies enrolling 4,418 patients.

As shown in Figure 2, the pooled proportion of mortality for patients without anticancer treatment was 0.97 (95%CI: 0.96 to 0.99). There was a statistically significant heterogeneity among pooled cohort studies (I^2 =93%, P < 0.00001).

RCTs: Data on mortality was extractable from the control arm of all 15 RCTs (1,031 patients). Figure 2 shows that the pooled proportion of mortality for patients in the control arm (without active treatment) was 0.96 (95% CI: 0.94 to 0.98). There was a statistically significant heterogeneity among pooled control arm of RCTs (I^2 =80%, P < 0.00001).

Combined (Cohort and RCTs): Pooled proportion of mortality across the 22 studies was 0.97 (95%CI: 0.96 to 0.98). Because these two designs are inherently different from each other, we conducted separate analyses. However, as shown in Figure 2, test for subgroup differences showed no statistically significant heterogeneity between the two study designs (P = 0.28).

Insert **Figure 2** about here

Median Survival

Cohort: Data on median overall survival was extractable from six cohort studies (4,125 patients). As shown in Figure 3, the pooled mean survival was 11.94 months (95%CI: 10.07 to 13.8). There was a statistically significant heterogeneity among pooled cohort studies (I^2 =97%, P < 0.00001).

RCTs: Data on median overall survival was extractable from all 15 RCTs (1,031 patients). The pooled mean survival for patients in the control arm was 5.03 months (95% CI: 4.17 to 5.89)

(Figure 3). There was a statistically significant heterogeneity among pooled control arm of RCTs (I^2 =90%, P < 0.00001).

Combined (Cohort and RCTs): Pooled proportion of mean survival across the 21 studies was 7.15 months (95%CI: 5.87 to 8.42). Test for subgroup differences showed statistically significant heterogeneity between the two study designs (I^2 = 97.7, P < 0.00001). Thus, the mean survival was influenced by study design (Figure 3).

Insert **Figure 3** about here

Sensitivity analysis

To assess the robustness of overall results according to the study design (cohort vs. RCT) as well as explore the reasons for observed heterogeneity in the pooled proportion of mortality and mean survival, we conducted additional sensitivity analyses. For both cohort studies and RCTs, we conducted sensitivity analyses according to methodological quality criteria, funding source, and study location. For RCTs only, we conducted additional sensitivity analyses according to type of control. The results of sensitivity analyses are summarized in Figure 4. Overall, the results remained unchanged in the sensitivity analyses. There were no statistically significant differences in the proportion of mortality.

Cohort: In cohort studies, there was no statistically significant difference in the proportion of mortality according to any methodological criteria of reporting. With respect to study location, the pooled proportion of mortality in cohort studies conducted at multicenter national locations was 0.95 (95%CI: 0.89 to 1.01) and at single institution was 0.98 (95%CI: 0.95 to 1.01) whereas

the pooled proportion of mortality in cohort studies conducted at unspecified locations was 0.87 (95%CI: 0.82 to 0.93). Test for overall interaction among these subgroups was statistically significant (P = 0.007). Regarding funding source, the pooled proportion of mortality in public-funded, unspecified funding sources, and public/industry-funded cohort studies were 1.00 (95%CI: 1.00 to 1.00), 1.00 (95%CI: 0.99 to 1.00), and 0.97 (95%CI: 0.96 to 0.98), respectively. The test for overall interaction among these subgroups was statistically significant (P < 0.0001).

RCTs: There was no statistically significant difference in the proportion of mortality according to methodological criteria of reporting, study location, and funding source. With respect to type of control, the pooled proportion of mortality in RCTs involving best supportive care, no treatment, placebo, supportive care, and symptomatic treatment as control were 0.90 (95%CI: 0.83 to 0.97) and in RCTs involving supportive care as control was 0.96 (95%CI: 0.92 to 1.00), 0.86 (95%CI: 0.81 to 0.92), 1.00 (95%CI: 0.99 to 1.01), 0.96 (95%CI: 0.92 to 1.00), and 0.97 (95%CI: 0.92 to 1.03), respectively. Test for overall interaction among these subgroups was statistically significant (P < 0.00001).

Insert **Figure 4** about here

We considered performing subgroup analysis based on median followup. However, only one cohort study [32] and five RCTs [38, 39, 41, 42, 45] reported these data. The median followup in the cohort study was 40 months whereas in the RCTs, the median followup was 2.7, 13, 26, 40, and 40 months, respectively. Given that survival of patients with cancer differs by stage, we considered performing analysis by cancer stage (I, II, III, vs. IV). However, only two cohort

studies (29%) and two RCTs (13%) reported data by stage. Thus, it was not possible to perform meta-analysis based on the four stages.

Discussion

This is the first study to provide most comprehensive data related to survival of lung cancer. The results show that prognosis of patients with lung cancer not receiving treatment is very high. Regardles of the study design (i.e. cohort versus RCTs) the findings were similar and did not differ according to disease severity. For example, all cohort studies assessed mortality in patients with early stage NSCLC (stage I/II) and all RCTs enrolled patients with advance stage NSCLC (stage III/IV). However, the mortality rates from cohort and RCTs essentially remained unchanged (97% vs 96%). Overall, included studies were of moderate methodological quality.

The findings from our study is similar to the study by Detterbeck and Gibson[4] which showed a 98% 5-year mortality rate for stage I/II lung cancer (median survival = 10 months). Despite the obvious similarity in results our study is significantly different in the conduct and analysis. For example, the study by Detterbeck and Gibson[4] did not employ a systematic approach to data collection and analysis (i.e. not a systematic review) and therefore the findings are not reproducible. The similarity in findings might be an artifact of play of chance. Furthermore, quantitative synthesis of results across included studies was not performed in the study by Detterbeck and Gibson[4] which was undertaken in our study. Another unique feature of our study lies in the inclusion of RCTs in addition to retrospective studies. None of the previous studies on the topic have utilized the approach of pooling data from one arm of RCTs for accurate assessment of prognosis. Therefore, due to the reasons enumerated here the study presented here is the most comprehensive to date reporting survival of NSCLC patients without

treatment.

Our study has some limitations. For example, we observed a statistically significant heterogeneity in pooled results which we could not explain through subgroup analyses. We suspect that the observed heterogeneity is clinical and not methodological. Specifically in the case of RCTs, the constitution of control arm varied across pooled studies. For example, five RCTs employed best supportive care as control, four had supportive care, two had placebo, two had no treatment and another two had symptomatic treatment as control. While, the definitions are very clear on placebo and no treatment, which was also explained by the sensitivity analyses (I^2 =0% for both subgroups), the composition of best supportive care, supportive care, and symptomatic treatment varied significantly across pooled studies. In these cases, the observed heterogeneity remained unexplained. Also, whereas a significant number of studies (11 of 15 RCTs) included had some form of treatment even if used for the purpose of symptom palliation, we were unable to assess the effect of the supportive treatment on survival based on available data. Thus, the clinical heterogeneity may be attributed to cancer stage of disease and/or differential therapies. Studies included had different followup period, however, due to limited data reported, we were unable to perform subgroup analysis based on median followup. How much this difference accounts for results is thus not known. It is also unclear whether results would have changed had we performed the analysis by cancer stages (I, II, III, vs. IV) as opposed to by stage I/II and III/IV. The former was not possible due to limited data reported. Because studies included enrolled patients with NSCLC, our results may not entirely apply to all lung cancer patients. However, it is important to note that a systematic review is limited by the

availability of data and we did include all available data related to prognosis of NSCLC patients without treatment.

Conclusion

The aim of this review was to estimate overall survival (natural history) in lung cancer when no anticancer therapy is provided. Our study shows that untreated lung cancer patients live on average for 7.15 months (95%CI: 5.87 to 8.42). Comprehensive data on the natural history of lung cancer is required for informed decision making by patients, physicians and researchers. For patients, it serves as the basis for their expected outcome with and without treatment, which is critical in cases of diseases with high mortality. For physicians, accurate and reliable information facilitates shared decision making with patients related to choice of interventions or no intervention. Most importantly, the findings are needed by researchers to avoid optimism bias.[8] A study by Djulbegovic et al. [8] assessed the role of optimism bias in a cohort of trials conducted by the National Cancer Institute Cooperative Groups and concluded that the optimism bias is the primary reason for inconclusive findings in the context of RCTs. Similarly, a systematic review by Gan and colleagues[9] showed that investigators tend to make overly optimistic assumptions regarding treatment benefits when designing RCTs. Accordingly, the results from our study will help researchers determine the most optimal rate of expected improvement in mortality with innovative/newer treatments.

Abbreviations

NSCLC = Non-Small Cell Lung Cancer; PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses; RCT = Randomized Controlled Trial

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HW and RM were responsible for the study's conception and design. HW acquired, analyzed and interpreted the data. HW drafted the manuscript while all other authors revised it critically for important intellectual content before giving approval of the final version to be published.

Acknowledgements

Supported by the US DoA grant #W81 XWH 09-2-0175 (PI: Djulbegovic).

References

1. Jemal, A., *Global Cancer Statistics (vol 61, pg 69, 2011).* Ca-a Cancer Journal for Clinicians, 2011. **61**(2): p. 134-134.
2. Ferlay, J., et al., *Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008.* International Journal of Cancer, 2010. **127**(12): p. 2893-2917.
3. Society, A.C., *Cancer Facts and Figures 2011*, in *America Cancer Society*2011.
4. Detterbeck, F.C. and C.J. Gibson, *Turning gray: the natural history of lung cancer over time.* J Thorac Oncol, 2008. **3**(7): p. 781-92.
5. Moons, K.G.M., et al., *Prognosis and prognostic research: what, why, and how?*
6. Christakis, N.A. and E.B. Lamont, *Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study.* Bmj, 2000. **320**(7233): p. 469-72.
7. Mackillop, W.J. and C.F. Quirt, *Measuring the accuracy of prognostic judgments in oncology.* J Clin Epidemiol, 1997. **50**(1): p. 21-9.
8. Djulbegovic, B., et al., *Optimism bias leads to inconclusive results-an empirical study.* Journal of Clinical Epidemiology, 2011. **64**(6): p. 583-93.
9. Gan, H.K., et al., *Assumptions of expected benefits in randomized phase III trials evaluating systemic treatments for cancer.* Journal of the National Cancer Institute, 2012. **104**(8): p. 590-8.
10. Daugherty, C.K. and F.J. Hlubocky, *What Are Terminally Ill Cancer Patients Told About Their Expected Deaths? A Study of Cancer Physicians' Self-Reports of Prognosis Disclosure.* Journal of Clinical Oncology, 2008. **26**(36): p. 5988-5993.
11. Moher, D., et al., *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.* PLoS medicine, 2009. **6**(7): p. e1000097.
12. Hemingway, H., R.D. Riley, and D.G. Altman, *Ten steps towards improving prognosis research.*
13. Hemingway, H., *Prognosis research: why is Dr. Lydgate still waiting?* J Clin Epidemiol, 2006. **59**(12): p. 1229-38.
14. Moons, K.G., et al., *Prognosis and prognostic research: what, why, and how?* Bmj, 2009. **338**: p. b375.
15. Chalmers, I., et al., *Table: Steps in finding evidence ("Level") for different types of question.*
16. Wilczynski, N.L. and R.B. Haynes, *Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey.* BMC Med, 2004. **2**: p. 23.
17. Haynes, R.B., et al., *Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey.* Bmj, 2005. **330**(7501): p. 1179.
18. Hayden, J.A., P. Cote, and C. Bombardier, *Evaluation of the quality of prognosis studies in systematic reviews.* Ann Intern Med, 2006. **144**(6): p. 427-37.
19. Laupacis, A., et al., *Users Guides to the Medical Literature .5. How to Use an Article About Prognosis.* Jama-Journal of the American Medical Association, 1994. **272**(3): p. 234-237.
20. *Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 [updated September 2009]. The Cochrane Collaboration, 2009. Available from www.cochrane-handbook.org.*

21. Hudak, P.L., D.C. Cole, and J.W. Frank, *Perspectives on prognosis of soft tissue musculoskeletal disorders.* Int J Rehabil Res, 1998. **21**(1): p. 29-40.
22. Altman, D.G., *Systematic reviews of evaluations of prognostic variables.* Bmj, 2001. **323**(7306): p. 224-8.
23. Stuart, A., *Kendall's advanced theory of statistics.* 1994. **102**(1): p. 173-80.
24. Hozo, S.P., B. Djulbegovic, and I. Hozo, *Estimating the mean and variance from the median, range, and the size of a sample.* BMC medical research methodology, 2005. **5**: p. 13.
25. Higgins, J.P.T., et al., *Measuring inconsistency in meta-analyses*, 2003. p. 557-560.
26. The Cochrane Collaboration, *Review Manager (RevMan)* 2008, The Nordic Cochrane Centre: Copenhagen.
27. Chadha, A.S., et al., *Survival in untreated early stage non-small cell lung cancer.* Anticancer Res, 2005. **25**(5): p. 3517-20.
28. Henschke, C.I., et al., *Small stage I cancers of the lung: genuineness and curability.* Lung Cancer, 2003. **39**(3): p. 327-30.
29. Hyde, L., et al., *Natural course of inoperable lung cancer.* Chest, 1973. **64**(3): p. 309-12.
30. McGarry, R.C., et al., *Observation-only management of early stage, medically inoperable lung cancer: poor outcome.* Chest, 2002. **121**(4): p. 1155-8.
31. Raz, D.J., et al., *Natural history of stage I non-small cell lung cancer: implications for early detection.* Chest, 2007. **132**(1): p. 193-9.
32. Vrdoljak, E., et al., *Survival analysis of untreated patients with non-small-cell lung cancer.* Chest, 1994. **106**(6): p. 1797-800.
33. Wisnivesky, J.P. and E.A. Halm, *Sex differences in lung cancer survival: do tumors behave differently in elderly women?* Journal of Clinical Oncology, 2007. **25**(13): p. 1705-12.
34. Anderson, H., et al., *Gemcitabine plus best supportive care (BSC) vs BSC in inoperable non-small cell lung cancer - a randomized trial with quality of life as the primary outcome.* Br J Cancer, 2000. **83**(4): p. 447-453.
35. Cartei, G., et al., *Cisplatin-cyclophosphamide-mitomycin combination chemotherapy with supportive care versus supportive care alone for treatment of metastatic non-small-cell lung cancer.* J Natl Cancer Inst, 1993. **85**(10): p. 794-800.
36. Cellerino, R., et al., *A randomized trial of alternating chemotherapy versus best supportive care in advanced non-small-cell lung cancer*, 1991. p. 1453-1461.
37. Cormier, Y., et al., *Benefits of polychemotherapy in advanced non-small-cell bronchogenic carcinoma.* Cancer, 1982. **50**(5): p. 845-9.
38. Cullen, M.H., et al., *Mitomycin, Ifosfamide, and Cisplatin in Unresectable Non–Small-Cell Lung Cancer: Effects on Survival and Quality of Life*, 1999. p. 3188-3194.
39. Elderly Lung Cancer Vinorelbine Italian Study Group, T., *Effects of Vinorelbine on Quality of Life and Survival of Elderly Patients With Advanced Non-Small-Cell Lung Cancer*, 1999. p. 66-72.
40. Ganz, P.A., et al., *Supportive care versus supportive care and combination chemotherapy in metastatic non-small cell lung cancer. Does chemotherapy make a difference?* Cancer, 1989. **63**(7): p. 1271-8.
41. Goss, G., et al., *Randomized Phase II Study of Gefitinib Compared With Placebo in Chemotherapy-Naive Patients With Advanced Nonâ€"Small-Cell Lung Cancer and Poor Performance Status*, 2009. p. 2253-2260.

42. Helsing, M., et al., *Quality of life and survival in patients with advanced non-small cell lung cancer receiving supportive care plus chemotherapy with carboplatin and etoposide or supportive care only. A multicentre randomised phase III trial.* European Journal of Cancer, 1998. **34**(7): p. 1036-1044.
43. Kaasa, S., et al., *Symptomatic treatment versus combination chemotherapy for patients with extensive non-small cell lung cancer.* Cancer, 1991. **67**(10): p. 2443-7.
44. Laing, A.H., et al., *TREATMENT OF INOPERABLE CARCINOMA OF BRONCHUS.* The Lancet, 1975. **306**(7946): p. 1161-1164.
45. Leung, W.T., et al., *Combined chemotherapy and radiotherapy versus best supportive care in the treatment of inoperable non-small-cell lung cancer.* Oncology, 1992. **49**(5): p. 321-6.
46. Quoix, E., et al., *[Is chemotherapy with cisplatin useful in non small cell bronchial cancer at staging IV? Results of a randomized study].* Bull Cancer, 1991. **78**(4): p. 341-6.
47. Rapp, E., et al., *Chemotherapy can prolong survival in patients with advanced non-small-cell lung cancer--report of a Canadian multicenter randomized trial*, 1988. p. 633-641.
48. Thongprasert, S., et al., *Relationship between quality of life and clinical outcomes in advanced non-small cell lung cancer: best supportive care (BSC) versus BSC plus chemotherapy.* Lung Cancer, 1999. **24**(1): p. 17-24.
49. Anderson, G. and H. Payne, *Response rate and toxicity of etoposide (VP-16) in squamous carcinoma of the lung: report from the Lung Cancer Treatment Study Group.* Semin Oncol, 1985. **12**(1 Suppl 2): p. 21-2.
50. Kuijpers, T., et al., *Systematic review of prognostic cohort studies on shoulder disorders.* Pain, 2004. **109**(3): p. 420-31.
51. Nijrolder, I., H. van der Horst, and D. van der Windt, *Prognosis of fatigue. A systematic review.* J Psychosom Res, 2008. **64**(4): p. 335-49.
52. Stang, A., *Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses.* Eur J Epidemiol, 2010. **25**(9): p. 603-5.
53. Cole, M.G., et al., *Persistent delirium in older hospital patients: a systematic review of frequency and prognosis.* Age Ageing, 2009. **38**(1): p. 19-26.
54. Hermens, M.L., et al., *The prognosis of minor depression in the general population: a systematic review.* Gen Hosp Psychiatry, 2004. **26**(6): p. 453-62.

Table 1 Characteristics of Studies Included in the Review

Study	N	Study period (years)	Disease Stage		Histology			Male	Median Age (years)
			I	II	squamous	adeno	large-cell		
(a) Cohort studies									
Raz 2007	1432	13	1432	NR	460	419	89	747	74
Wisnivesky 2007†	2344	8	NR	NR	NR	NR	NR	1292	NR
Chadha 2005	39	11	23	13	18	88	5	4	77
Henschke 2003	131	7	131	NR	NR	NR	NR	NR	NR
McGarry 2002†	49	5	NR	NR	NR	NR	NR	49	NR
Vrdoljak 1994	130	7	55	56	61	35	34	120	60
Hyde 1973	293	8	NR	NR	NR	NR	NR	NR	NR
Total/[Range]	4418	[5-13]	1641	68	539	542	128	2211	
(b) RCTs			**III**	**IV**					
Goss 2009[m]	101	2 [0.23]	17	84	25	46	11	61	76
Anderson 2000	150	2	92	58	NR	NR	NR	91	64
ELVIS 1999 [m]	78	1 [1.08]	22	56	33	29	3	69	74*
Cullen 1999 [m]	176	8 [2.17]	88	88	103	42	6	122	64
Thongprasert 1999	98	4	49	49	31	49	12	NR	60
Helsing 1998 [m]	26	5 [3.33]	3	23	5	17	4	18	65
Cartei 1993	50	7	NR	50	25	17	8	36	57
Leung 1992 [m]	66	4 [3.58]	58	NR	31	18	7	48	62
Cellerino 1991	61	3	61	NR	38	18	5	59	62
Quoix 1991	22	3	NR	22	NR	NR	NR	NR	NR
Kaasa 1991	43	3	NR	43	16	16	11	31	62*
Ganz 1989	26	2	NR	26	9	17	NR	23	NR
Rapp 1988	50	3	50	NR	12	24	12	38	58
Cormier 1982	17	2	17	NR	8	2	6	16	60
Laing 1975	67	2	15	20	23	5	9	59	64
Total/[Range]	1031	[1-8]	472	519	359	300	94	671	[57-76]

Note: **N** = Sample size or number of participants enrolled; **NR**= data not reported; † = Sample includes stage I and II cancer; **adeno** = adenocarcinoma; **squamous** = squamous cell carcinoma; **large-cell** = large-cell carcinoma; *=we recorded mean age where median age was not reported or not extractable, [m] = median follow-up in parenthesis

Table 2 Methodological Quality of Lung Cancer Prognosis Studies

	Study Design/Domain/Criterion	Criteria fulfilled	
		n/N	%
Cohort studies (11 items)			
	Participation bias		
A	Population of interest is adequately described for key characteristics[15]	7/7	100
B	Study setting and geographic location is adequately described[15]	7/7	100
C	Baseline sample is adequately described for key characteristics[15]	4/7	57
D	Inclusion and exclusion criteria are adequately described[15]	5/7	71
E	There is adequate participation in the study by all eligible patients[15]	7/7	100
	Attrition bias		
F	Follow-up is sufficiently long for outcome to occur (≥ 6 months)[16,18,19,46]	6/7	86
G	Patients with missing data were reported[15,17]	7/7	100
	Outcome measurement		
H	Definition of outcome is provided *a priori* [15]	7/7	100
I	Objective definition of outcome is provided[15,16,18,19]	7/7	100
	Data analysis and reporting		
J	Alpha error and/or beta error is specified *a priori*	2/7	29
K	Frequencies of most important data (e.g., outcomes) are presented[18,19,47]	7/7	100
Randomized Controlled Trials (15 items)			
	Participation bias		
L	Population of interest is adequately described for key characteristics[15]	15/15	100
M	Study setting and geographic location is adequately described[15]	7/15	47
N	Baseline sample is adequately described for key characteristics[15]	14/15	93
O	Inclusion and exclusion criteria are adequately described[15]	14/15	93
P	Patients were balanced in all aspects except the intervention	15/15	93
	Attrition bias		
Q	Follow-up is sufficiently long for outcome to occur (≥ 6 months)[16,18,19,46,48]	8/15	53
R	Proportion of sample completing the study is adequate (≥80%)[15,16,18,47,49,50]	9/15	60
S	Description of withdrawal (incomplete outcome data) is provided[15,17]	15/15	100
T	Characteristics of dropouts versus completers is provided[15]	2/15	13
	Outcome measurement		
U	Definition of outcome is provided *a priori*[15]	15/15	100
V	Objective definition of outcome is provided[15,16,18,19]	15/15	100
	Data analysis and reporting		
W	Alpha error and/or beta error is specified *a priori*	7/15	47
X	Data analysis was based on intention to treat analysis principle[17]	9/15	53
Y	Frequencies of most important data (e.g., outcomes) are presented[18,19,47]	15/15	100

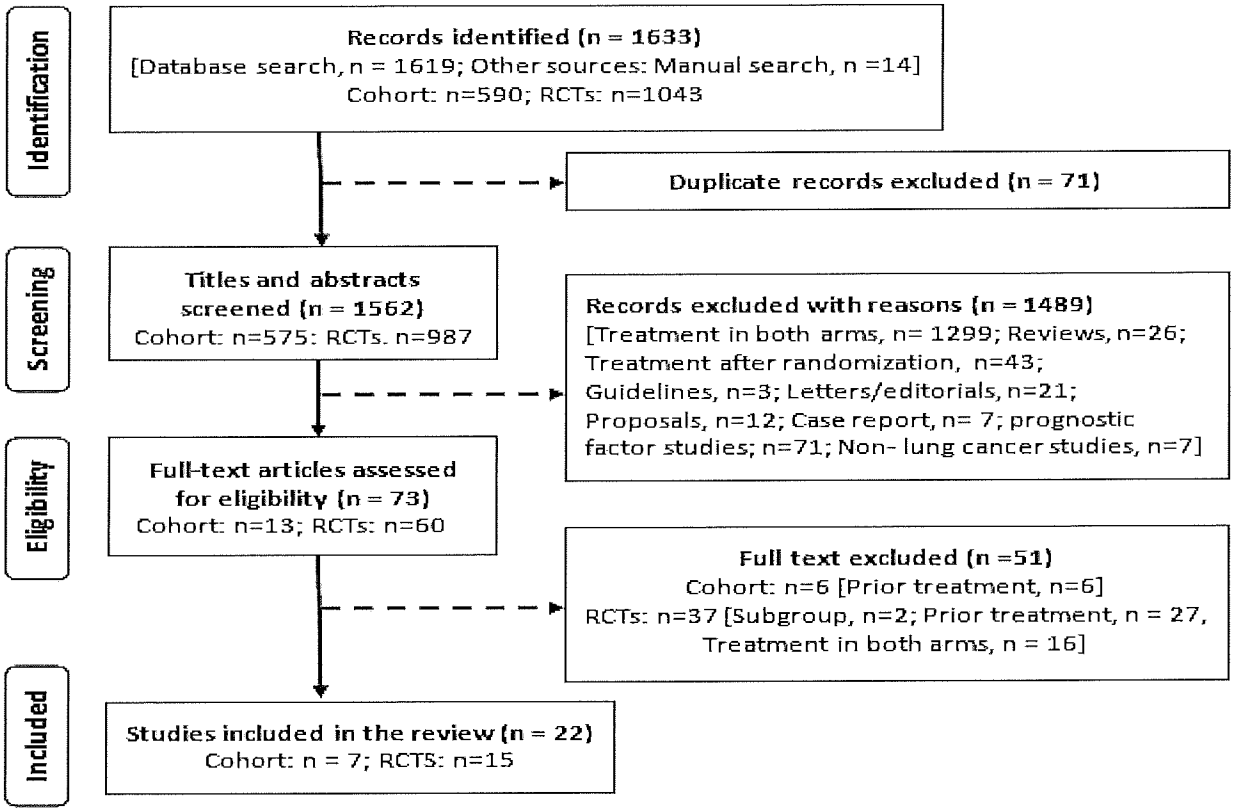


Figure 1. A flow diagram depicting the literature search process.

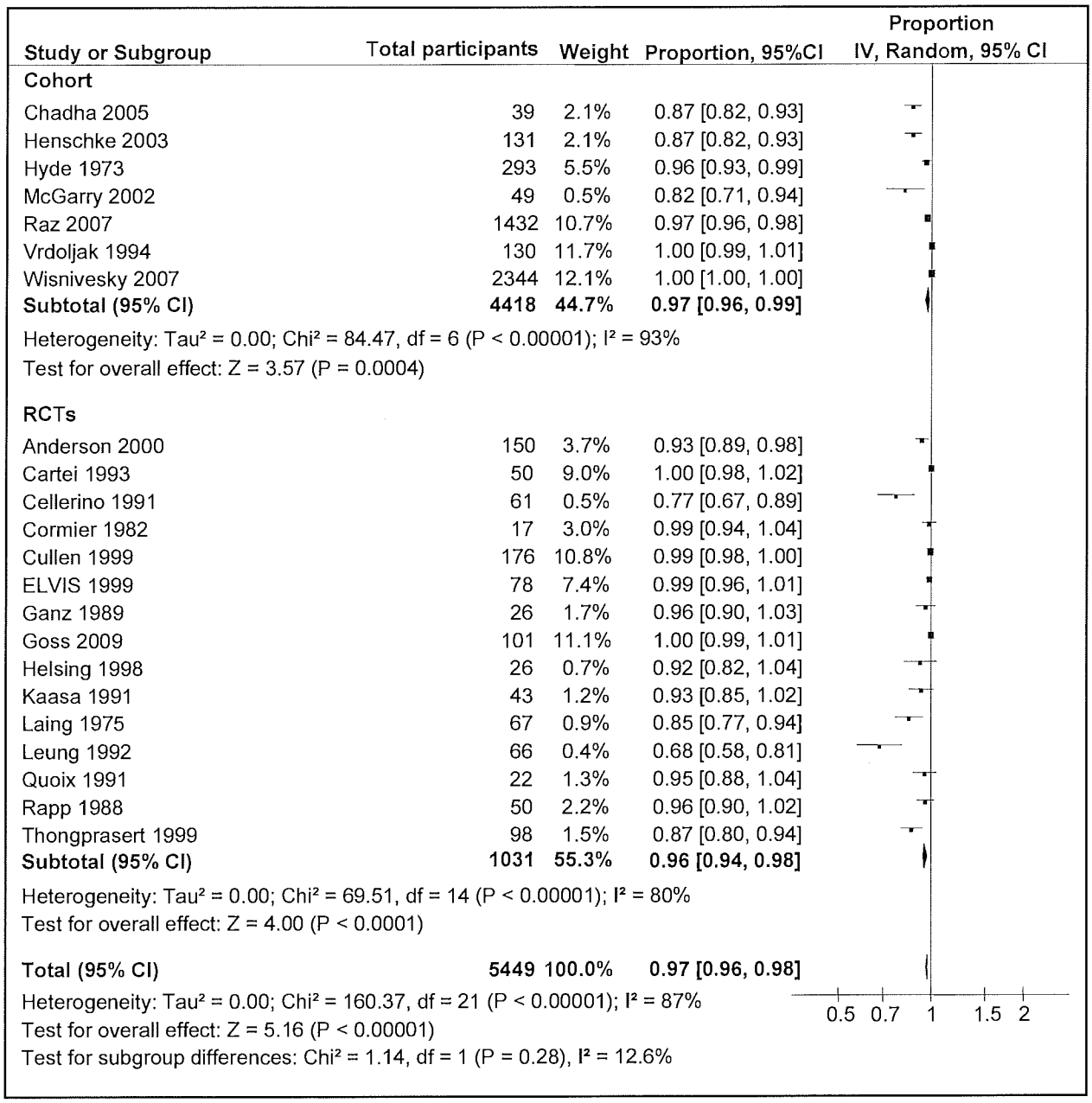


Figure 2. Pooled proportion of mortality in lung cancer studies. The size of each square is proportional to the weight of the study (inverse variance).

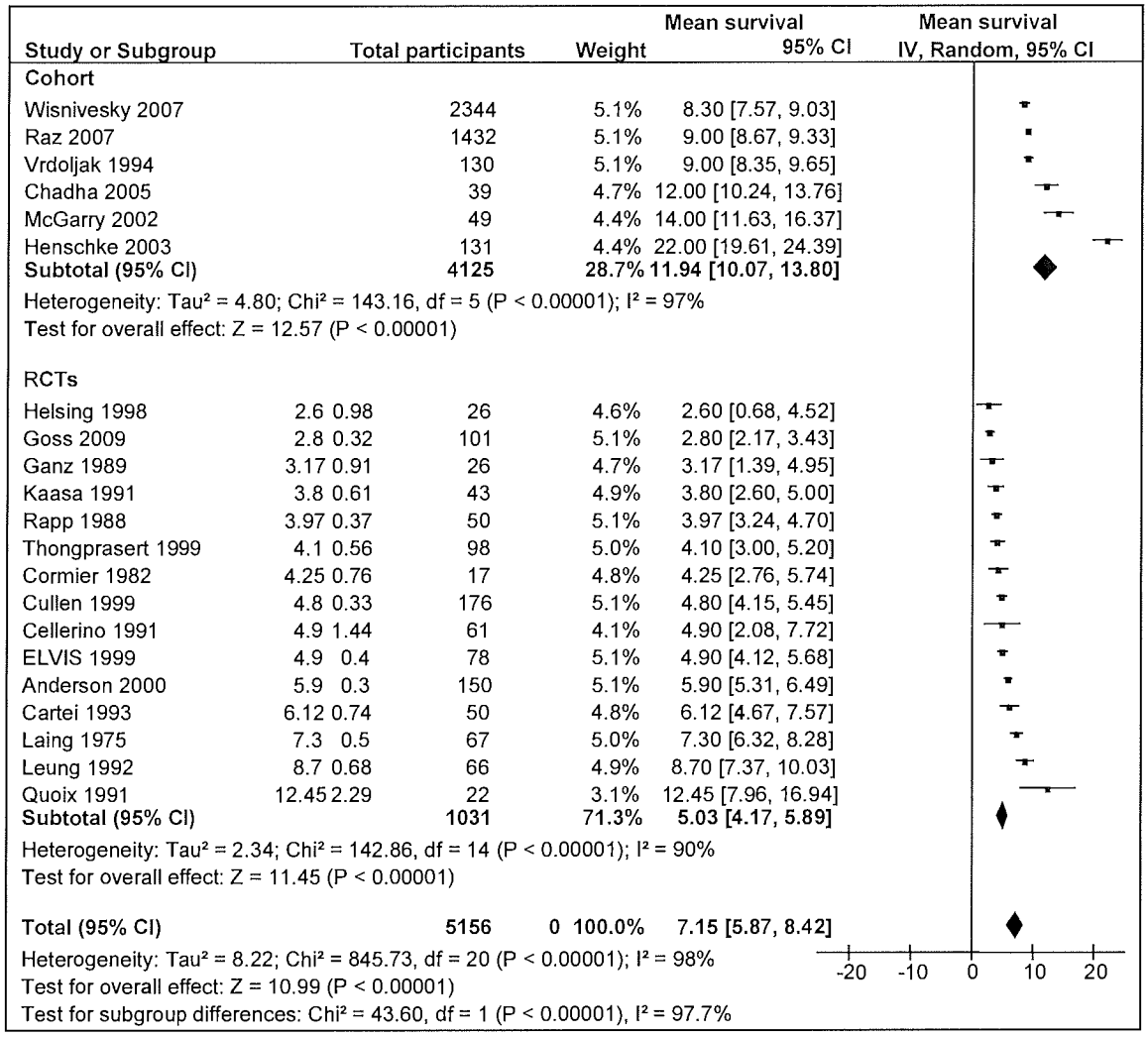


Figure 3. Pooled mean survival and heterogeneity between subgroups. The size of each square is proportional to the weight of the study (inverse variance).

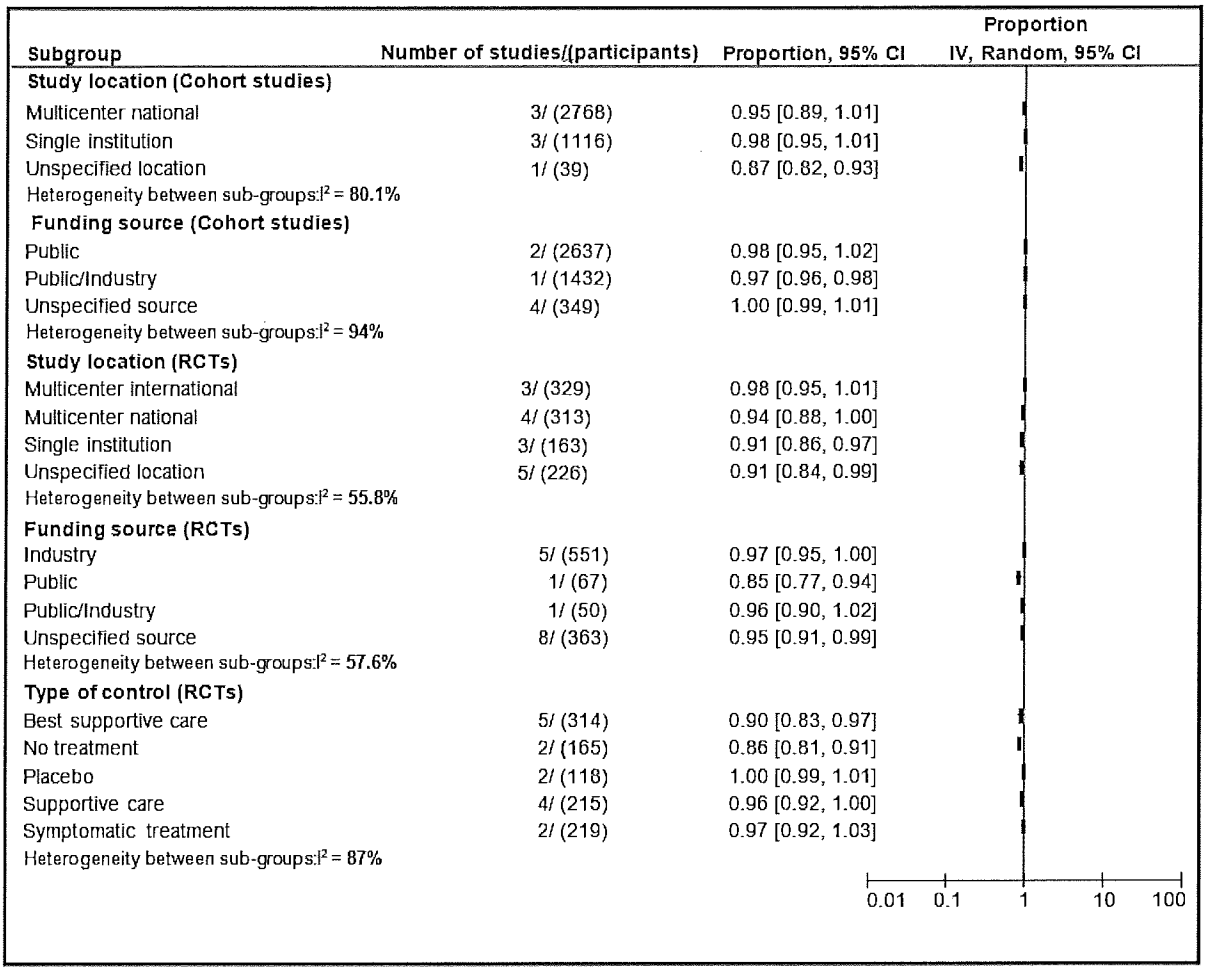


Figure 4. Pooled proportions of mortality and heterogeneity between subgroups. The size of each square is proportional to the weight of the study (inverse variance).

PLOS | ONE

A Flexible Alternative to the Cox Proportional Hazards Model for Assessing the Prognostic Accuracy of Hospice Patient Survival

Branko Miladinovic[1]*, Ambuj Kumar[1], Rahul Mhaskar[1], Sehwan Kim[2], Ronald Schonwetter[2], Benjamin Djulbegovic[1,3]

1 Center for Evidence Based Medicine and Health Outcomes Research, University of South Florida, Tampa, Florida, United States of America, **2** HPC Healthcare, Temple Terrace, Florida, United States of America, **3** H. Lee Moffitt Cancer Center & Research Institute, Tampa, Florida, United States of America

Abstract

Prognostic models are often used to estimate the length of patient survival. The Cox proportional hazards model has traditionally been applied to assess the accuracy of prognostic models. However, it may be suboptimal due to the inflexibility to model the baseline survival function and when the proportional hazards assumption is violated. The aim of this study was to use internal validation to compare the predictive power of a flexible Royston-Parmar family of survival functions with the Cox proportional hazards model. We applied the Palliative Performance Scale on a dataset of 590 hospice patients at the time of hospice admission. The retrospective data were obtained from the Lifepath Hospice and Palliative Care center in Hillsborough County, Florida, USA. The criteria used to evaluate and compare the models' predictive performance were the explained variation statistic R^2, scaled Brier score, and the discrimination slope. The explained variation statistic demonstrated that overall the Royston-Parmar family of survival functions provided a better fit ($R^2 = 0.298$; 95% CI: 0.236–0.358) than the Cox model ($R^2 = 0.156$; 95% CI: 0.111–0.203). The scaled Brier scores and discrimination slopes were consistently higher under the Royston-Parmar model. Researchers involved in prognosticating patient survival are encouraged to consider the Royston-Parmar model as an alternative to Cox.

Citation: Miladinovic B, Kumar A, Mhaskar R, Kim S, Schonwetter R, et al. (2012) A Flexible Alternative to the Cox Proportional Hazards Model for Assessing the Prognostic Accuracy of Hospice Patient Survival. PLoS ONE 7(10): e47804. doi:10.1371/journal.pone.0047804

Editor: Raya Khanin, Memorial Sloan Kettering Cancer Center, United States of America

Received May 11, 2012; **Accepted** September 21, 2012; **Published** October 17, 2012

Funding: This study was supported by the United States Army Medical Research and Material Command grant DOA W81 XWH-09-0175. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bmiladin@health.usf.edu

Introduction

Prognostic models are often used to estimate the length of patient survival and improvement in the accuracy of prognosis translates into superior quality of patient care. Precise prognosis of survival using modeling techniques requires rigorous methods for the development and testing of the accuracy of prognostic models. Developing a prognostic model entails having accurate patient data for prognosis, and selecting clinically relevant candidate predictors and measures of model performance, usually in the context of multivariable regression [1]. This process produces patient performance scores that allow for classification of patients into different risk groups [2,3,4].

In the hospice setting, accurate prognostication of survival affords patients and their families a vital opportunity to attend to matters such as planning, prioritizing, and preparing for death [5]. Predicting patient survival is a complex decision making process involving numerous subjective and numerical factors that have substantial variation which may lead to poor prediction of life expectancy. Many physicians practice optimism or avoidance, thus overestimating survival at times by a factor of five [6]. Implementing appropriate statistical methodologies translates into improved accuracy of prognosis and superior quality of care. Predictions based on appropriate statistical modeling have been shown to be superior to physicians' prognostication [4,7].

The Cox proportional hazards (CPH) model [8] is the most commonly-used survival prediction model [4,9]. In the hospice and palliative settings, demographic and clinical covariates are often included in CPH to predict patient survival [10,11]. The appeal of the model is its analytic simplicity and that the baseline survival function does not need to be defined *apriori*–it is absorbed when the likelihood function is maximized (note that "baseline" refers to zero values of the covariates, not to time equal to zero). It is possible to estimate the baseline survival function for the CPH model conditional on the estimated regression coefficients. However, this is highly rigid as the smoothing of the underlying function depends on the proportional hazards assumption, which may not be supported by the data and is often overlooked by the investigators [9]. Essentially, the CPH model was designed to measure the effects of covariates on the changing hazard function and not to model patient survival. A flexible family of functions which allows for parametrically modeling the baseline survival function is more appropriate, especially if the proportional hazards assumption is violated in the CPH [12]. The baseline survival has for the most part been ignored because it is left undefined in the CPH model.

In this manuscript we compare CPH with an alternative method of estimating survival in the form of the class of flexible Royston-Parmar (RP) parametric functions [12]. We use the Palliative Performance Scale (PPS) [13] from a cohort of hospice patients. Results from systematic reviews have shown that the patient PPS score is an accurate measure of patient survival in the palliative setting [7,11]. Furthermore, PPS and CPH model have been used to construct meaningful hospice patient survival estimates in the form of a life expectancy table and survival nomogram [14] and to validate prognosticating scales for hospice patient survival [15,16,17].

In addition to PPS, other risk factors such as age, cancer status and gender have been reported to be significant predictors of palliative patient survival in several studies [11,14]._ENREF_18 In our study we did not adjust for other risk factors because though they may be significant predictors of survival for the cohort of patients in our dataset, they may not be in other palliative settings. Our goal was to demonstrate that the RP family of parametric functions allowed for a direct and flexible modeling of the baseline survival and that it might be formulated so that the impact of the proportional hazard assumption is minimized. We determined if the overall performance and discriminatory ability of RP family of parametric functions is superior to CPH in the sample by using models that were derived and tested on the whole dataset (naïve validation) and using (internal) cross-validation. It is important to note that the RP parametric functions have not been applied to prognostic models in the hospice and palliative settings. It is also important to note that we did not perform external validation, which entailed using a different data set than the one used to create the model[3]. In the next section we briefly discuss PPS, introduce the statistical models and measures of model performance.

Methods

Study sample and palliative performance

The patient data were obtained from the Lifepath Hospice and Palliative Care Center licensed since 1983 to serve in Hillsborough County, Florida. Hospice care focuses on pain control and symptom management. To avoid selection bias, we retrospectively and sequentially extracted data for 590 patients who, as of January 2009 were deceased. This study was a retrospective review of the deceased patients' medical records. Only data pertaining to outcomes were collected; personal information was not collected and the data were de-identified prior to analysis. Since we did not collect any information that can identify deceased patients or their family members, under HIPPA rules and regulations (45 CFR 164.512) the requirement for consent does not apply. The study and consent procedures were approved by the University of South Florida Institutional Review Board prior to study initiation. Two research assistants extracted all data necessary to populate the model variables and two faculty members randomly checked 25% of the data for accuracy. The models were tested against observed survival duration.

The Palliative Performance Scale (PPS) was developed and reported by Anderson et al. [13] as a measure of palliative patients' functional status. The scale has 11 possible mutually exclusive levels, which are based on five domains: six levels of ambulation, six levels of activity and evidence of disease, five levels of self-care, five levels of food intake and four levels of consciousness. The scale ranges from PPS of 0% (deceased patient) to PPS of 100% (ambulatory and healthy patient). Numerous studies have studied its prognostic accuracy of survival in a variety of settings and found it provides meaningful estimates of patient survival [10,14,15,18,19,20,21,22,23]. PPS has been found to be both valid and reliable [24].

Model selection and validation

Validating a prognostic model is generally accepted to mean that given a patient population it works in a data set other than the one it is applied to[2,25]. In other words, the model needs to be tested using a different data set than the one used to create the model[3]. It is also generally accepted that the validation process should follow guidelines and that un-validated prognostic models should not be applied in clinical practice [3,4,26]. When validating a prognostic survival model in the regression framework, most attention has been on the value of the prognostic index based on covariates, while the role of the baseline survival function has been largely ignored.

The role of the baseline survival is significant as it quantifies the absolute patient survival probabilities over time. For a vector of covariates $\mathbf{x}$ and parameter vector $\boldsymbol{\beta}$, the survival function $S(t; \mathbf{x})$ at

Table 1. Patient characteristics.

Variable	Result
Total no. of patients	590 (100%)
Age at Treatment	
<45	37 (6.3%)
45–64	187 (31.7%)
65–74	110 (18.6%)
75–84	129 (21.9%)
85+	127 (21.5%)
Gender	
Male	293 (49.7%)
Female	295 (50%)
Unknown	2 (0.3%)
No. of patients with cancer/noncancer	
Noncancer	363 (61.5%)
Cancer	227 (38.5%)
Diagnosis category for cancer	
Brain	10 (1.7%)
Gastrointestinal	35 (5.9%)
Genital-female	12 (2%)
Genital-male	12 (2%)
Head and neck	8 (1.4%)
Hematopoietic	10 (1.7%)
Pancreas	24 (4.2%)
Respiratory	49 (8.3%)
Skin	2 (0.3%)
Urinary	4 (0.6%)
Other	61 (10.3%)
Diagnosis category for noncancer	
AIDS	12 (2%)
Cardiovascular	74 (12.5%)
Neurological	119 (20.2%)
Respiratory	37 (6.3%)
Other	121 (20.6%)

doi:10.1371/journal.pone.0047804.t001

Table 2. Survival time by age, gender, diagnosis and initial PPS.

	Survival Times (in Days)			
Variable	**Mean (95% CI)**	**Median (95% CI)**	**Range**	**No. of Patients (%)**
Total no. of patients				
Overall	14 (12, 17)	6 (5, 6)	1–371	590
Age at Treatment				
<45	15 (8, 22)	8 (4,12)	1–95	37 (6.3%)
45–64	14 (11, 17)	7 (5, 9)	1–114	187 (31.7%)
65–74	14 (8, 20)	5 (4, 6)	1–271	110 (18.6%)
75–84	14 (8, 20)	6 (5, 7)	1–371	129 (21.9%)
85+	15 (9, 21)	5 (4, 6)	1–313	127 (21.5%)
Gender				
Male	14 (10, 18)	6 (5, 7)	1–371	293 (49.7%)
Female	15 (11, 19)	6 (5, 7)	1–271	295 (50%)
No. of patients with cancer				
Noncancer	12 (8, 16)	5 (4, 6)	1–371	363 (61.5%)
Cancer	17 (14, 20)	9 (7, 11)	1–113	227 (38.5%)
Diagnosis category for cancer				
Brain	27 (16, 39)	28 (14, 42)	3–55	10 (1.7%)
Gastrointestinal	21 (14, 29)	11 (5,17)	1–82	35 (5.9%)
Genital-female	15 (6, 24)	8 (1, 15)	2–55	12 (2%)
Genital-male	26 (7, 45)	13 (4, 22)	1–100	12 (2%)
Head and neck	10 (2, 18)	5 (1, 9)	1–36	8 (1.4%)
Hematopoietic	4 (2, 6)	3 (1, 5)	1–10	10 (1.7%)
Pancreas	18 (7, 29)	7 (3, 11)	1–113	24 (4.2%)
Respiratory	15 (10, 20)	10 (7, 13)	1–71	49 (8.3%)
Skin	11	11	11–11	2 (0.3%)
Urinary	25 (1, 58)	9 (1, 39)	4–76	4 (0.6%)
Other	17 (12, 22)	9 (5, 12)	1–103	61 (10.3%)
Diagnosis category for noncancer				
AIDS	18 (3, 33)	8 (1, 15)	1–85	12 (2%)
Cardiovascular	14 (5, 23)	5 (3, 7)	1–271	74 (12.5%)
Neurological	8 (5, 11)	5 (4,6)	1–77	119 (20.2%)
Respiratory	25 (1, 49)	3 (1, 5)	1–371	37 (6.3%)
Other	11 (1, 15)	5 (4, 6)	1–174	121 (20.6%)
Initial PPS Score				
PPS 10%	5 (3, 7)	3 (2, 4)	1–77	188 (32.6%)
PPS 20%	16 (8, 24)	5 (4, 6)	1–371	125 (21.7%)
PPS 30%	15 (11, 19)	7 (5, 9)	1–140	123 (21.4%)
PPS 40%	24 (18, 30)	14 (11, 17)	1–147	96 (16.7%)
PPS 50–80%	28 (21, 35)	18 (9, 27)	1–76	44 (7.6%)

doi:10.1371/journal.pone.0047804.t002

time t for the CPH model is commonly expressed as $S(t;x)=[S_0(t)]^{\exp(x\beta)}$, where S_0 (t) is the baseline survival function, i.e. survival function when all the covariates **x** are equal to zero. In the CPH framework, the estimation of the prognostic index $\mathbf{x}\boldsymbol{\beta}$ does not require the formulation of the baseline cumulative survival function S_0 (t), which itself can be estimated conditional on the covariate estimates. The two popular methods for estimating baseline survival S_0 (t) are the Breslow and Kalbfleisch-Prentice methods [27]. Both give similar results in practice, but can lead to "choppy" estimates of the baseline function and are dependent on the proportional hazards assumption.

When the goal of a survival analysis is to estimate hazard ratios (the effect of covariates on the changing hazard function), the baseline function is of no consequence. The CPH is appropriate as the baseline function gets absorbed when coefficient $\boldsymbol{\beta}$s are estimates by the method of partial log likelihood. However, when the goal is to prognosticate patient survival, there is a need for more flexibility in modeling the baseline survival.

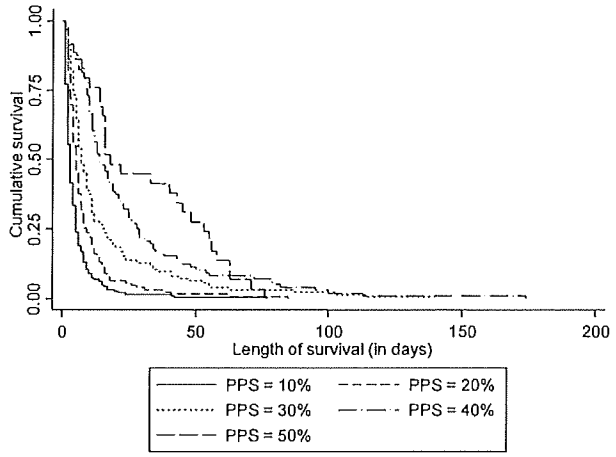


Figure 1. Kaplan-Meier survival curves by initial PPS.
doi:10.1371/journal.pone.0047804.g001

An alternative to the CPH is the RP family of models that resembles the generalized linear models and can be viewed as a parametric extension Cox proportional hazard models [12]. The models are framed to rely on the transformation g(.), such that $g(S(t;x))=g(S_0(t))+x\beta$. The transformation g(.) can be either from the proportional hazard, proportional odds, Aranda-Ordaz or probit families [12]. We did not consider the Aranda-Ordaz family in this study due to possible interpretational difficulties [12]. Under the proportional hazard link function, the hazard ratio estimates are nearly identical to those estimated under CPH. The attractive feature of the RP baseline survival function is that its shape is preserved, but the location of the baseline distribution function can vary, which allows for flexible model recalibration. Also, the estimate $g(s_0(t))$ is implemented on log-time scale. It is generally gently curved and smooth, making survival estimates more accurate.

In the RP framework, if the proportional hazard assumption is violated, the probit-link function $g(s)=-\Phi^{-1}(s)$ can be applied, where $\Phi^{-1}(.)$ is the inverse standard normal distribution function. The baseline survival function $s_0(t)$ is approximated and smoothed by a restricted cubic spline function with m interior knots. Splines are piecewise polynomials that ensure the overall curve is smooth (see Royston and Parmar [12] for details). Spline-based survival models such as RP have been empirically shown to be superior when the proportional hazard assumption is violated [28]. The optimal number of knots and the comparison among different RP models can be found using the minimum combination of Akaike Information Criterion (AIC), Bayes Information Criterion (BIC) and explained variation statistic R^2 [29,30]. The AIC is defined in the usual manner as - 2Log(likelihood) + 2(No. of model parameters), while BIC equals - 2Log(likelihood) + (No. of model parameters)*Log(n). In survival analysis n is interpreted as the number of events rather than the number of patients. The placement of knots in spline modeling is an issue. We have placed the knots at equally spaced centiles of the log-survival times, following published recommendations [31]. For example, for m = 1 the knot is at the 50^{th} centile, for m = 2 the knots are at the 33^{th} and 67^{th} centiles, etc.

We compared RP and CPH by performing internal validation (assessing validity in the population where the development data originated from) on the whole data set (naïve) and using split-sample cross-validation. We performed 10-fold cross-validation by splitting the data into development

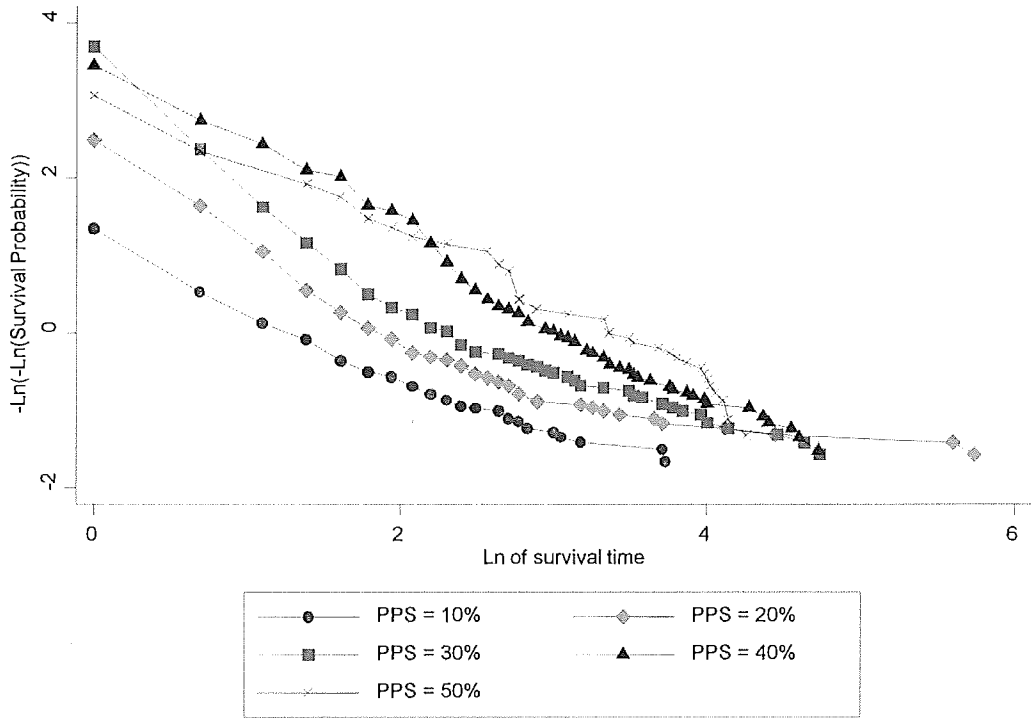


Figure 2. Test of the proportional hazards assumption under CPH for initial PPS.
doi:10.1371/journal.pone.0047804.g002

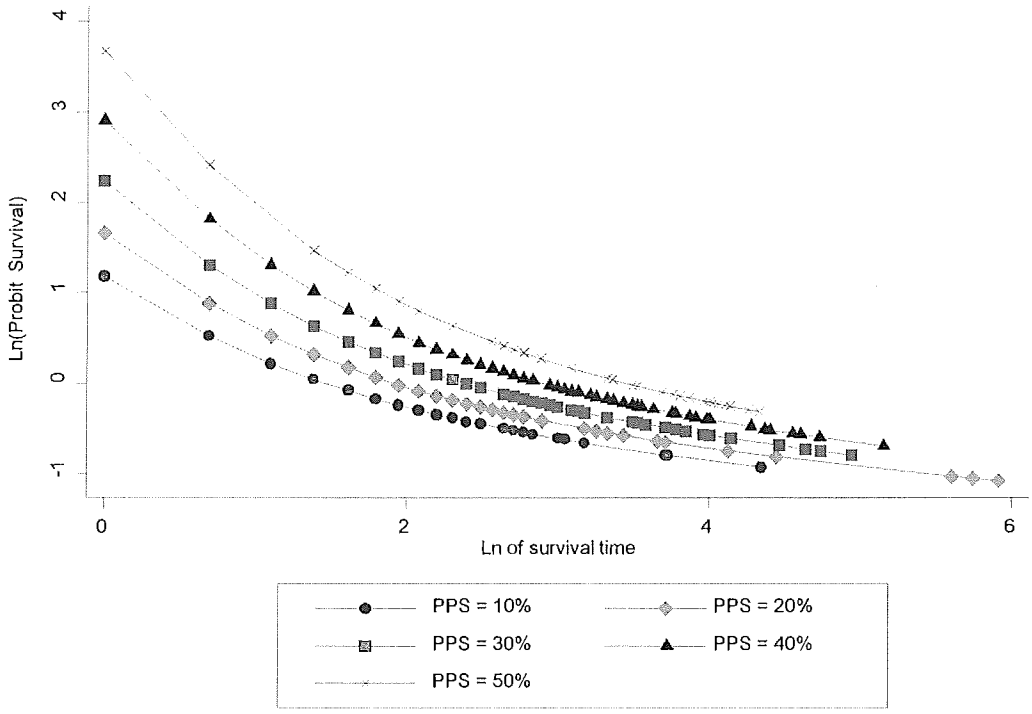


Figure 3. Test of the probit assumption under RP for initial PPS.
doi:10.1371/journal.pone.0047804.g003

and validation sets and repeating the process 20 times. The methods can be readily implemented in Stata [32,33] statistical software using the *stpm* [29] and *stpm2* [34] commands, or in open source statistical software R as *flexsurv* package [35].

Assessment of model performance

Model performance is the ability of the estimated risk score to predict survival and is assessed using the measures of explained variation, calibration, and discrimination. Calibration refers to how closely the predicted survival at a pre-specified time agrees with the observed survival. For cross-validation, we compared the average fitted probabilities of survival under RP and CPH for the first 15 days to observed probabilities estimated non-parametrically using Kaplan-Meier curves [36].

The Brier score is a quadratic scoring rule that calculates the differences between the actual outcomes and predicted probabilities[37]. Given the predicted probability of survival p_i at time t for patient i, and Y_i binary (0–1, dead-alive) variable, the Brier score is defined as $\sum(Y_i(1-p_i)^2+(1-Y_i)p_i^2)$. A Brier score of 0 indicates a perfect model, while 0.25 indicates a non-informative model (the value achieved when issuing a predicted probability of 50% to each patient). The Brier score may be scaled by its maximum $\text{Brier}_{\max}=(1 - \text{mean}(p_i))\ \text{mean}(p_i)$ to obtain $\text{Brier}_{\text{scaled}}=(1-\frac{\text{Brier}}{\text{Brier}_{\max}})100\%$. The scaled Brier scores range from 0% to 100% and have interpretation similar to the Pearson correlation coefficient[38].

Table 3. Criteria for the choice of scale in the RP model.

No. of knots m	PH	PO	Probit
	AIC, BIC, R^2	AIC, BIC, R^2	AIC, BIC, R^2
0	2033, 2042, 0.156	1887, 1896, 0.321	1872, 1881, 0.295
1	1889, 1902, 0.178	1883, 1896, 0.322	1858, 1871, 0.298
2	1871, 1888, 0.170	1870, 1887, 0.312	1857, 1874, 0.296
3	1870, 1892, 0.172	1870, 1892, 0.311	1858, 1880, 0.297
4	1865, 1892, 0.171	1865, 1891, 0.310	1855, 1881, 0.296
5	1866, 1896, 0.171	1865, 1896, 0.309	1856, 1886, 0.296

doi:10.1371/journal.pone.0047804.t003

For a particular risk score, discrimination is the ability to differentiate between the patients who died versus those who survived. The Kaplan-Meier plot of survival for patients in different risk groups can be used to test for separation, indicating that the different risk groups are well defined [39]. For a statistical model, the global measure of the model's discriminatory power is the explained variation statistic R^2, which measures the variation explained by the fitted model [40]. Higher values of R^2 indicate greater discrimination. In this study we implement R^2 for survival models, as described by Royston and Sauerbrei,[41].

The discrimination or Yates slope is a measure of how well the subjects with and without the outcome are separated. It is defined as the absolute difference in mean predictions of survival (mean[p_i]) between those who died and those who survived at time t[2]. The scaled Brier scores and discrimination slopes were calculated separately for the (naïve) model using the whole dataset and the model derived using cross-validation for t = 1, 2... 100 days. Higher scaled Brier scores and discrimination slopes represent better model performance.

All statistical calculation were performed using Stata version 11.2 [32,33].

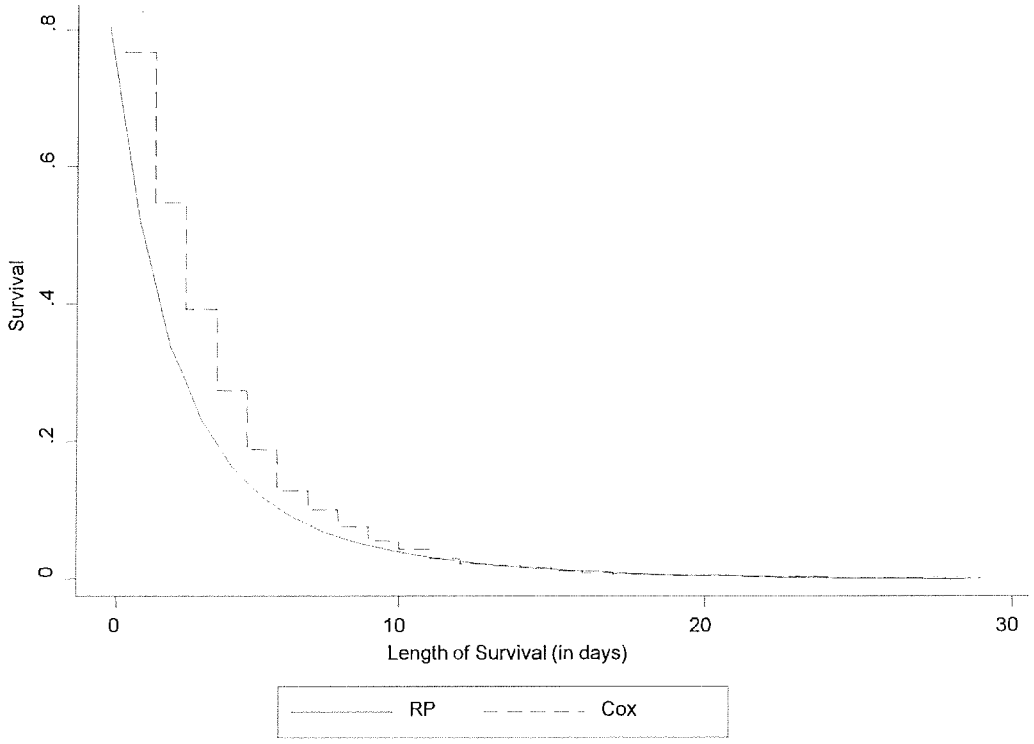


Figure 4. Baseline survival functions under CPH and RP models.
doi:10.1371/journal.pone.0047804.g004

Results

Description of the data source

The patient characteristics of the retrospective cohort are summarized in Table 1. The cohort consisted of 293 males (49.7%) and 295 females (50.0%), and 2 (0.3%) with unknown gender. The data were collected starting from patients' entry into hospice care until death for all 590 patients. The mean, median and range of survival times for the patients by PPS at admission, age, gender, cancer status, and diagnosis category are given in Table 2. The table shows that the median survival was fairly evenly distributed across age groups and gender, but unevenly across the cancer status and initial diagnosis category. All patients were assigned PPS at the time of

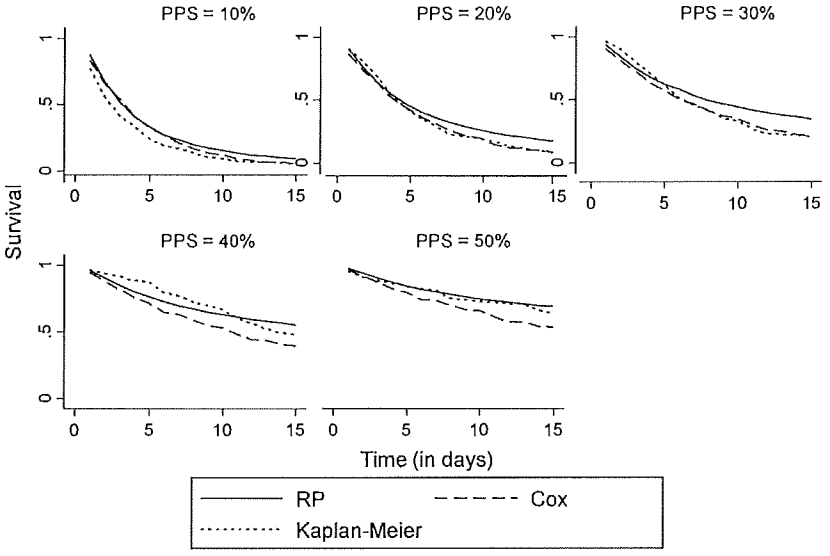


Figure 5. Predicted survival by PPS under RP and CPH compared with the Kaplan-Meier estimates in the validation data.
doi:10.1371/journal.pone.0047804.g005

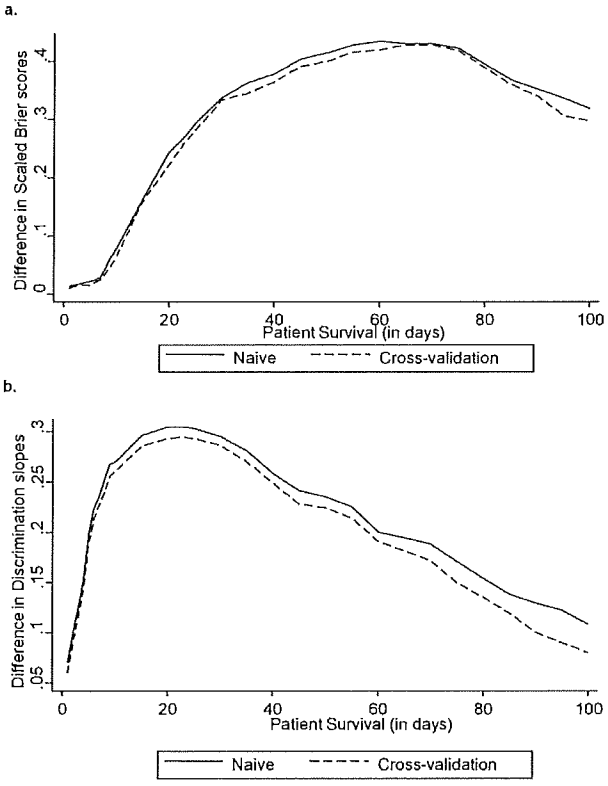


Figure 6. Difference between Brier scores for RP and CPH models (6a) and between discrimination slopes for RP and CPH models (6b) as a function of patient survival times in the naïve (whole) data set and cross-validated data set. Both are consistently higher for RP indicating better accuracy and discrimination.
doi:10.1371/journal.pone.0047804.g006

admission to hospice care. Since PPS score of 0% means that the patient is dead, the data were transformed so that the PPS score of 10% was set as the baseline. There were only 15 total observations for PPS = 60%, 70%, 80%, so they were combined with PPS = 50% to obtain meaningful survival estimates. Fourteen patients had missing values for PPS.

The time of admission was the starting point for survival time. The Kaplan-Meier curves stratified by initial PPS level are shown in Figure 1. The curves show good separation indicating that the different risk groups are well defined. The log-rank test for equality of survival curves was highly significant at $P = 0.001$. The global test based on Schoenfeld residuals showed that the proportional hazard assumption was violated for PPS (P-value <0.001), which can also be seen from the un-parallel natural log-plot of survival curves (Figure 2).

Table 3 lists AIC, BIC and R^2 values for 5 knots under the proportional hazard, proportional odds and probit RP families; the minimum combination in each is underlined. The number of optimal knots was found to be $m = 1$ under the probit model. The improvement in fit with the probit model can be seen from the parallel survival curves of log-probit against natural log time (Figure 3).

R^2 was higher in the RP model ($R^2 = 0.298$; 95% CI: 0.236–0.358) than the Cox model ($R^2 = 0.156$; 95% CI: 0.111–0.203), indicating that the RP model explained significantly more variation than CPH. To illustrate the differences for the baseline function, Figure 4 shows plots of the CPH and RP baseline survival functions. The CPH baseline survival is "choppy" to approximately day 12, while the RP is smooth. The two baseline functions converged at around day 12.

Cross-validation showed that the relation between the two predicted survival estimates is approximately linear, with RP model consistently estimating a higher probability, which is particularly evident for higher scores of PPS corresponding to longer survival times (Figure 5). Overall, the predicted probabilities under RP tended to be closer to the Kaplan-Meier estimates than CPH. The plot of the consistently positive differences between RP and CPH scaled Brier scores (Figure 6a) and discrimination slopes (Figure 6b) showed that the RP model discriminated better across patient survival times for both the full (naïve) and cross-validated models. This suggested that the higher value of R^2 under RP was not due to over-fitting.

Discussion

The results from our study show that RP family of models predicts survival more accurately than CPH through its flexible modeling of the baseline survival function. Using the RP flexible baseline function modeling would allow for more precise calibration in the prognostication phase than CPH. As Figure 5 illustrates, the predicted RP survival probabilities are consistently higher for higher values of PPS, and closer to the Kaplan-Meier estimates of survival. We suspect that both the robust modeling of baseline survival and overall model fit provide for better survival estimation.

There are limitations to our study, the primary one being the use of retrospective data. The RP family of parametric functions needs to be applied prospectively to assess accuracy of prognostic models through external validation. Furthermore, the dataset was limited to the hospice setting with no censored observations and with majority of patients having a very short follow-up time. For future studies, application of the proposed methodology should account for these limitations, and comparisons with parametric prognostic survival models should be explored.

The flexible models discussed in this paper could greatly improve the ability of researchers to accurately predict survival. An advantage of RP is that it can be used to validate published models for which the original individual patient data are unavailable. If the scale used (hazard, probit or odds), the knot positions, and the estimates of prognostic indices are known, then it would be possible to use RP. In the case of CPH this is not possible, since the baseline function would not be available.

Acknowledgments

The authors wish to thank Dr. Jane Carver for her help in preparing the manuscript.

Author Contributions

Conceived and designed the experiments: BM BD. Analyzed the data: BM. Contributed reagents/materials/analysis tools: RS SK. Wrote the paper: BM BD AK RM.

References

1. Royston P, Moons KG, Altman DG, Vergouwe Y (2009) Prognosis and prognostic research: Developing a prognostic model. BMJ 338: b604.

2. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, et al. (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21: 128–138.

3. Vickers AJ (2011) Prediction models: revolutionary in principle, but do they do more good than harm? J Clin Oncol 29: 2951–2952.
4. Moons KG, Altman DG, Vergouwe Y, Royston P (2009) Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 338: b606.
5. Steinhauser KE, Christakis NA, Clipp EC, McNeilly M, McIntyre L, et al. (2000) Factors considered important at the end of life by patients, family, physicians, and other care providers. JAMA 284: 2476–2482.
6. Christakis NA, Lamont EB (2000) Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. BMJ 320: 469–472.
7. Chow E, Harth T, Hruby G, Finkelstein J, Wu J, et al. (2001) How accurate are physicians' clinical predictions of survival and the available prognostic tools in estimating survival times in terminally ill cancer patients? A systematic review. Clin Oncol (R Coll Radiol) 13: 209–218.
8. Cox DR, Oakes D (1984) Analysis of survival data. London;New York: Chapman and Hall. viii, 201 p.p.
9. Mallett S, Royston P, Waters R, Dutton S, Altman DG (2010) Reporting performance of prognostic models in cancer: a review. BMC Med 8: 21.
10. Downing M, Lau F, Lesperance M, Karlson N, Shaw J, et al. (2007) Meta-analysis of survival prediction with Palliative Performance Scale. J Palliat Care 23: 245–252; discussion 252–244.
11. Lau F, Clontier-Fisher D, Kuziemsky C, Black F, Downing M, et al. (2007) A systematic review of prognostic tools for estimating survival time in palliative care. J Palliat Care 23: 93–112.
12. Royston P, Parmar MK (2002) Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med 21: 2175–2197.
13. Anderson F, Downing GM, Hill J, Casorso L, Lerch N (1996) Palliative performance scale (PPS): a new tool. J Palliat Care 12: 5–11.
14. Lau F, Downing M, Lesperance M, Karlson N, Kuziemsky C, et al. (2009) Using the Palliative Performance Scale to provide meaningful survival estimates. J Pain Symptom Manage 38: 134–144.
15. Harrold J, Rickerson E, Carroll JT, McGrath J, Morales K, et al. (2005) Is the palliative performance scale a useful predictor of mortality in a heterogeneous hospice population? J Palliat Med 8: 503–509.
16. Olajide O, Hanson L, Usher BM, Qaqish BF, Schwartz R, et al. (2007) Validation of the palliative performance scale in the acute tertiary care hospital setting. J Palliat Med 10: 111–117.
17. Head B, Ritchie CS, Smoot TM (2005) Prognostication in hospice care: can the palliative performance scale help? J Palliat Med 8: 492–502.
18. Fainsinger RL, Demoissac D, Cole J, Mead-Wood K, Lee E (2000) Home versus hospice inpatient care: discharge characteristics of palliative care patients in an acute care hospital. J Palliat Care 16: 29–34.
19. Morita T, Tsunoda J, Inoue S, Chihara S (2001) Effects of high dose opioids and sedatives on survival in terminally ill cancer patients. J Pain Symptom Manage 21: 282–289.
20. Virik K, Glare P (2002) Validation of the palliative performance scale for inpatients admitted to a palliative care unit in Sydney, Australia. J Pain Symptom Manage 23: 455–457.
21. Morita T, Tsunoda J, Inoue S, Chihara S (1999) Validity of the palliative performance scale from a survival perspective. J Pain Symptom Manage 18: 2–3.
22. Lau F, Bell H, Dean M, Downing M, Lesperance M (2008) Use of the Palliative Performance Scale in survival prediction for terminally ill patients in Western Newfoundland, Canada. J Palliat Care 24: 282–284.
23. Lau F, Maida V, Downing M, Lesperance M, Karlson N, et al. (2009) Use of the Palliative Performance Scale (PPS) for end-of-life prognostication in a palliative medicine consultation service. J Pain Symptom Manage 37: 965–972.
24. Ho F, Lau F, Downing MG, Lesperance M (2008) A reliability and validity study of the Palliative Performance Scale. BMC Palliat Care 7: 10.
25. Altman DG, Royston P (2000) What do we mean by validating a prognostic model? Stat Med 19: 453–473.
26. Altman DG, Vergouwe Y, Royston P, Moons KG (2009) Prognosis and prognostic research: validating a prognostic model. BMJ 338: b605.
27. Kalbfleisch JD, Prentice RL (2002)The statistical analysis of failure time data. Hoboken, N.J.: J. Wiley.. xiii, 439 p.p.
28. Binquet C, Abrahamowicz M, Mahboubi A, Jooste V, Faivre J, et al. (2008) Empirical study of the dependence of the results of multivariable flexible survival analyses on model selection strategy. Stat Med 27: 6470–6488.
29. Royston P (2001) Flexible parametric alternatives to the Cox model, and more. Stata Journal 1: 1–28.
30. Schwarz G (1978) Estimating Dimension of a Model. Annals of Statistics 6: 461–464.
31. Durrleman S, Simon R (1989) Flexible regression models with cubic splines. Stat Med 8: 551–561.
32. Stata Version 11 [computer program]. 9 ed. College Station, TX: Stata Corporation; 2010.
33. Royston P (2011) Flexible parametric survival analysis using stata : beyond the Cox model. College Station, TX: Stata Press.
34. Lambert PC, Royston P (2009) Further development of flexible parametric models for survival analysis. The Stata Journal 9: 265–290.
35. Jackson C (2012) Flexible parametric survival models.
36. Cook NR, Buring JE, Ridker PM (2006) The effect of including C-reactive protein in cardiovascular risk prediction models for women. Ann Intern Med 145: 21–29.
37. Graf E, Schmoor C, Sauerbrei W, Schumacher M (1999) Assessment and comparison of prognostic classification schemes for survival data. Stat Med 18: 2529–2545.
38. Hu B, Palta M, Shao J (2006) Properties of R(2) statistics for logistic regression. Stat Med 25: 1383–1395.
39. Altman DG (2009) Prognostic models: a methodological framework and review of models for breast cancer. Cancer Invest 27: 235–243.
40. Royston P (2006) Explained variation for survival models. The Stata Journal 6: 1–14.
41. Royston P, Sauerbrei W (2004) A new measure of prognostic separation in survival data. Stat Med 23: 723–748.

34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Towards a Classification Model to Identify Hospice Candidates in Terminally Ill Patients

Eleazar Gil-Herrera, Ali Yalcin, Athanasios Tsalatsanis, Laura E. Barnes and Benjamin Djulbegovic

Abstract— **This paper presents a Rough Set Theory (RST) based classification model to identify hospice candidates within a group of terminally ill patients. Hospice care considerations are particularly valuable for terminally ill patients since they enable patients and their families to initiate end-of-life discussions and choose the most desired management strategy for the remainder of their lives. Unlike traditional data mining methodologies, our approach seeks to identify subgroups of patients possessing common characteristics that distinguish them from other subgroups in the dataset. Thus, heterogeneity in the data set is captured before the classification model is built. Object related reducts are used to obtain the minimum set of attributes that describe each subgroup existing in the dataset. As a result, a collection of decision rules is derived for classifying new patients based on the subgroup to which they belong. Results show improvements in the classification accuracy compared to a traditional RST methodology, in which patient diversity is not considered. We envision our work as a part of a comprehensive decision support system designed to facilitate end-of-life care decisions. Retrospective data from 9105 patients is used to demonstrate the design and implementation details of the classification model.**

I. Introduction

A. Hospice referral criteria

Hospice is designed to provide comfort and support to terminally ill patients and their families. According to Medicare regulations, a patient should be referred to hospice if his/her life expectancy is approximately 6 months or less [1]. However, most patients are not referred to hospice in a timely manner [2, 3] and therefore they do not reap the well-documented benefits of hospice services. A premature hospice referral translates to a patient losing the opportunity to receive potentially effective treatment, which may prolong their life. Conversely, a late hospice referral may deprive patients and their families of enjoying the benefits offered. Therefore, accurate prognostication of life expectancy is of vital importance for terminal patients as well as for their families and physicians.

B. Prognostic models for estimating survival of terminally ill patients

Survival prognostic models range from traditional statistical and probabilistic techniques [4-10], to models based on artificial intelligence such as neural networks [11, 12], decision trees [13, 14] and rough set methods [15, 16]. The primary goal of survival prognostic models is to provide accurate information regarding life expectancy and/or determine the association between prognostic factors and survival. Typically, the information derived by prognostic models is presented in terms of probability of death within a time period. Recent systematic reviews [17, 18] have highlighted the necessity of prediction models that can be easily integrated into clinical practice and facilitate end-of-life clinical decision-making.

Several important issues demand particular consideration when developing clinical classification models: First, clinical data, representing patient records that include symptoms and clinical signs, are not always well defined and are represented with *vagueness* [19]. Therefore, it is very difficult to classify cases in which small differences in the value of an attribute may completely change the classification of a patient and, as a result, the treatment decisions [20]. Second, clinical data may present *inconsistencies*, which means that it is possible to have more than one patient with the same description but with different outcomes. Third, the results of prognostic models should be readily interpretable to enable practical and posteriori inspection and interpretation by the treating physician or an expert system [21]. Finally, prognostic models should consider the heterogeneity in clinical data, i.e. the existence of patient diversity presented in terms of risk of disease and responsiveness to treatment [22, 23]. This consideration will enable a prognostic model to identify possible subgroups of patients for which certain covariates do not influence their classification. The practical implications of such considerations are associated with the ability to customize the prognostic model for each subgroup of patients (e.g. expensive and/or potentially harmful tests may be avoided for particular subgroups).

Rough Set Theory (RST) [24], a mathematical tool for representing and reasoning about vagueness and inconsistency in data sets, has been used in a number of applications dealing with modeling medical prognosis [15, 16, 25-28]. For example, Tsumoto et al. [25], provide a framework to model medical diagnosis rules showing theoretically that the characteristics of medical reasoning reflect the concepts of approximation established in RST. Komorowski et al. [26], show that RST is useful to extract medical diagnosis rules to identify a group of patients for whom performing a test that is costly or invasive is redundant or superfluous in the prognosis of a particular medical condition. Recently, [28] highlighted features of RST for integrating into medical applications. For example, RST has the ability to handle imprecise and uncertain information and provides a schematic approach for analyzing data without initial assumptions on data distribution.

This work was supported in part by the Department of Army under grant #W81 XWH-09-2-0175.

E. Gil-Herrera and A. Yalcin are with the Department of Industrial and Management System Engineering, University of South Florida, Tampa, FL 33620, USA(e-mail: eleazar@mail.usf.edu, ayalcin@eng.usf.edu).

A. Tsalatsanis, L. Barnes and B. Djulbegovic are with the Center for Evidence Based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL 33612, USA (e-mail: atsalats@health.usf.edu, lbarnes@health.usf.edu, bdjulbeg@health.usf.edu)

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

In our previous work [29], we proposed the use of RST to predict the life expectancy of terminally ill patients using a *global reduction* [30] methodology to identify the most significant attributes for building the classification model. However, we found that the number of attributes used in the model was barely reduced and therefore produced long decision rules. Moreover, considering the number of discretization categories associated with each attribute, the generated decision rules were built to describe each object in the training set and therefore, they were poorly suited for classifying new cases.

Here, we propose the use of an alternative attribute reduction methodology that aims to identify groups of patients that share common characteristics that distinguish them from the rest of the patients. As a result, we obtain subgroups of patients from which different sets of significant attributes are identified. The decision rules generated in this manner contain fewer attributes and therefore are more suitable to classify new patients. Moreover, by studying each subgroup, we can reason about how a different rule-set is applied to a particular patient.

The rest of the paper describes details of the proposed RST based methodology to provide a classifier that properly discriminates patients into two groups: those who survive at least 180 days after evaluation for hospice referral and those who do not. ROSETTA [31] software is used to perform the analysis described in the remainder of the paper.

II. METHODOLOGY

A. Data Set

The dataset used in this study consists of the 9105 cases from the SUPPORT (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments) prognostic model dataset [30]. We consider all variables used in the SUPPORT prognostic model [3] as condition attributes, i.e. the 10 physiologic variables along with the diagnosis groups, age, number of days in the hospital before entering the study, presence of cancer, and neurologic function. Data collection and patient selection procedures are detailed in [3]. Attributes names and descriptions are listed in Table I. As the decision attribute, we define a binary variable (Yes/No) "deceases_in_6months" using the following two attributes from the SUPPORT prognosis model dataset:

• death: represents the event of death at any time up to NDI date (National Death Index date: Dec 31, 1994).

• D.time: number of days of follow up

The values of the decision attribute are calculated converting the "D.time" value in months and comparing against the attribute "death" as follows:

• If "D.time" < 6 months and "death" is equal to 1 (the patient died within 6 months) then "deceased_in_6months" is "Yes". Otherwise, it is implicit that a patient survived the 6-month period; hence, "deceased_in_6months" is "No".

B. Rough Set Theory Data Representation

Based on RST, the data set is represented as:

$$T = (U, A \cup \{d\}) \quad (1)$$

TABLE I. CONDITION ATTRIBUTES

Name	Description
alb	Serum albumin
bili	Bilirubin
crea	Serum creatinine
hrt	Heart rate
meanbp	Mean arterial blood pressure
pafi	Arterial blood gases
resp	Respiratory rate
sod	Sodium
temp	Temperature (Celsius)
wblc	White blood cell count
dzgroup	Diagnosis group
age	Patient's age
hday	Days in hospital at study admit
ca	Presence of cancer
scoma	SUPPORT coma score based on Glasgow coma scale

where T, represents the dataset in the form of a table. Each row represents an object and each column represents an attribute. U is a non-empty finite set of objects and the set A is a non-empty finite set of attributes called the condition attributes. In our case, an object designates a terminally ill patient and an attribute $a \in A$ designates each of the fifteen condition attributes that describe a patient (Table I). For every attribute, the function $a: U \rightarrow V_a$ makes a correspondence between an object in U to an attribute value V_a which is called the value set of a. The set T incorporates an additional attribute $\{d\}$ called the decision attribute. The system represented by this scheme is called a decision system.

C. Development of the Classification Model

This process typically involves numerous steps, such as data preprocessing, discretization, reduction of attributes, rule induction, classification and interpretation of the results. Details on the data preprocessing and data discretization for this data set are described in [29]. The ultimate goal of this process is to generate decision rules, which are used to classify each patient as surviving or not surviving within the defined period of time. A decision rule has the form: *if A then B* ($A \rightarrow B$), where A is called the condition and B the decision of the rule.

Here, we are focusing on an alternative method of reducing the attribute dimensions and identify different subgroups of similar patients in the data set. In [32], two types of reducts are defined:

1) Global Reducts:

Consists of the minimal set of attributes that preserve the structure of the entire data set. A set $B \subseteq A$ is called a global reduct if the indiscernibility relation using attributes B is equal to the indiscernibility relation using all the condition attributes A, i.e.:

$$IND(B) = IND(A), \text{ where,}$$

$$IND(B) = \{(u_i, u_j) \in U^2 : \forall a_k \in B, a_k(u_i) \neq a_k(u_j)\}$$

As an example, consider the following global reduct obtained from the data set containing 12 condition attributes:

G_RED = {age, dzgroup, scoma, ca, meanbp, wblc, hrt, resp, temp, bili, crea, sod}

Using *G_RED*, few patients will have exactly the same attribute-value combinations because the number of discretization categories associated with each attribute is high. Thus, the decision rules generated are too specific to the cases in the training set and therefore may not be able to classify new cases accurately. Moreover, the fact that global reducts represent the entire data set makes it difficult to detect the presence of heterogeneous groups in the data meaning that the causes of diversity between the patient outcomes will remain unknown.

2) Object related reducts (ORR):

Represents the minimal attribute subsets that discern an object $u \in U$ from the rest of objects belonging to a different decision class. Mathematically, an ORR $R_u \subseteq A$ is defined as:

$\forall u_i \in U : d(u_i) \neq d(u_j) \Rightarrow \exists a_k \in R_u : a_k(u_i) \neq a_k(u_j)$, *where* $u_i \neq u_j$.

An ORR is the minimal and vital information that is used to partition the universe of objects into smaller, homogeneous subgroups, where objects within a subgroup are related by means of information described by the ORR. Decision rules generated by this scheme will usually contain fewer attributes and are more suitable to classify new cases. Some decision rules contain a different set of attributes applicable for a particular subgroup of patients.

III. RESULTS

The two methods for dimensionality reduction produce a set of reducts. The number of reducts and decision rules obtained are presented in Table II. Based on the decision rules generated, patients are classified as surviving or not surviving the six-month period. A standard voting algorithm [30] is used for this purpose. Table III, presents the performance of two classification models based on each type of reduct generation described. The performance of each classification model is represented in terms of *sensitivity*, *specificity*, *Area under the Receiver Operating Characteristic curve* (AUC) and *coverage* of the model. A 5-fold cross validation procedure was applied to estimate the performance of each classification model, where, the entire data set is randomly divided into five subsets (folds). Then, each fold (20% of the data set) is used once as a testing set, while the remaining folds (80%) are used for training. The process is repeated five times and the results are averaged to provide an estimate for the classifier performance.

Compared to the Global reduct approach, the ORR approach has enhanced the classification performance in terms of AUC and sensitivity. Moreover the decision rules generated are able to classify all new cases.

TABLE II. NUMBER OF REDUCTS AND DECISION RULES GENERATED – GLOBAL VS. ORR

Method	Number of reducts	Number of rules
Global reducts	99	647,223
ORR	11,894	68,492

TABLE III. CLASIFICATION RESULTS – GLOBAL VS. ORR

Method	Sensitivity	Specificity	AUC	Coverage
Global reducts	73.67%	44.05%	61.8%	86.43%
ORR	86.92%	39.2%	71.9%	100%

IV. DISCUSSION

Analyzing the information obtained from the ORR, we can identify groups of patients for whom it is possible to evade costly, invasive or even unnecessary tests required by the prediction model. For example, the following two ORRs generate rules independent of the *Pafi* score (associated with the patient's blood gases), without reducing the classification accuracy. The importance of such finding becomes apparent considering that in clinical practice *Pafi* is not collected routinely for patients outside the Intensive Care Unit (ICU).

- ORR = {Age, dzgroup, meanbp} generates the following decision rules:
  - if age= [45, 60) AND dzgroup = (Lung Cancer) AND meanbp=[60, 70) then: Survive = 22.86%, Die = 77.14%.
  - if age= [45, 60) AND dzgroup = (CHF) AND meanbp=[100, 120) then: Survive = 82.93%, Die = 17.07%.
  - if age= [70, 75) AND dzgroup = (COPD) AND meanbp=[80,100) then: Survive = 84.21%, Die = 15.79%.
- ORR = {Age, dzgroup, hrt, crea} generates the following decision rules:
  - if age= [45, 60) AND dzgroup = (CHF) AND hrt=[100,110) and crea[1.95, *] then: Survive = 83.33%, Die = 16.67%.
  - if age= [75,85) AND dzgroup = (CHF) AND hrt=[50,110) and crea[0.5, 1.5) then: Survive = 82.19%, Die = 17.81%.

Consequently, the use of *Pafi* test in patients that belong to one of those groups defined by the ORR's will not improve the prognostication accuracy.

Our approach demonstrates features that make it particularly suitable for use in clinical decision-making. It is a patient-centric methodology which is able to predict without the use of unnecessary, expensive and/or invasive procedures for certain subgroups of patients. Consequently, selection of attributes upon which a decision is to be made is critical to minimizing healthcare costs and maximizing the quality of patient care. Finally, considering that more than one ORR could discern each patient, the information acquired offers several options dependent on the attribute values available for each individual patient.

V. FUTURE WORK

The number of ORR and the decision rules generated depends on the number of condition attributes and its categories. For clinical datasets, which contain large numbers of condition attributes, the number of ORRs and decision rules generated can be extremely large to be

evaluated directly by human experts. Therefore, the interpretation and analysis of the ORRs and their decision rules requires the use of a well-defined methodology.

Compared to our previous work [29], the results presented in this paper show an improvement in the classifier performance. However, further research need to be conducted in order to achieve a reliable prognostic model.

REFERENCES

[1] L.R. Aiken and NetLibrary Inc., "Dying, death, and bereavement," in Book Dying, death, and bereavement, *Series Dying, death, and bereavement*, 4th ed. Lawrence Erlbaum Associates, 2000.

[2] N.A. Christakis, "Timing of referral of terminally ill patients to an outpatient hospice.," *J Gen Intern Med*, vol. 9, (no. 6), pp. 314-20, Jun 1994.

[3] A. Tsalatsanis, L.E. Barnes, I. Hozo, and B. Djulbegovic, "Extensions to Regret-based Decision Curve Analysis: An application to hospice referral for terminal patients," *BMC Med Inform Decis Mak*, vol. 11, pp. 77, 2011.

[4] W.A. Knaus, D.P. Wagner, E.A. Draper, J.E. Zimmerman, M. Bergner, P.G. Bastos, C.A. Sirio, D.J. Murphy, T. Lotring, and A. Damiano, "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults," *Chest*, vol. 100, (no. 6), pp. 1619-1636, December, 1991.

[5] W.A. Knaus, F.E. Harrell, J. Lynn, L. Goldman, R.S. Phillips, A.F. Connors, N.V. Dawson, W.J. Fulkerson, R.M. Califf, N. Desbiens, P. Layde, R.K. Oye, P.E. Bellamy, R.B. Hakim, and D.P. Wagner, "The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults," *Annals of Internal Medicine*, vol. 122, (no. 3), pp. 191-203, February, 1995.

[6] D.W. Hosmer and S. Lemeshow, *Applied survival analysis regression modeling of time to event data*, New York, NY: Wiley, 1999.

[7] J.R. Beck, S.G. Pauker, J.E. Gottlieb, K. Klein, and J.P. Kassirer, "A convenient approximation of life expectancy (the "DEALE"): II. Use in medical decision-making," *The American Journal of Medicine*, vol. 73, (no. 6), pp. 889-897, 1982.

[8] I. Hyodo, T. Morita, I. Adachi, Y. Shima, A. Yoshizawa, and K. Hiraga, "Development of a Predicting Tool for Survival of Terminally Ill Cancer Patients," *Japanese Journal of Clinical Oncology*, vol. 40, (no. 5), pp. 442-448, May 1, 2010.

[9] D. Porock, D. Parker-Oliver, G. Petroski, and M. Rantz, "The MDS Mortality Risk Index: The evolution of a method for predicting 6-month mortality in nursing home residents," *BMC Research Notes*, vol. 3, (no. 1), pp. 200, 2010.

[10] P.K.J. Han, M. Lee, B.B. Reeve, A.B. Mariotto, Z. Wang, R.D. Hays, K.R. Yabroff, M. Topor, and E.J. Feuer, "Development of a Prognostic Model for Six-Month Mortality in Older Adults With Declining Health," *Journal of Pain and Symptom Management*, vol. 43, (no. 3), pp. 527-539, 2012.

[11] G. Clermont, D.C. Angus, S.M. DiRusso, M. Griffin, and W.T. Linde-Zwirble, "Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models," *Critical Care Medicine*, vol. 29, (no. 2), 2001.

[12] R. Dybowski, V. Gant, P. Weller, and R. Chang, "Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm," *The Lancet*, vol. 347, (no. 9009), pp. 1146-1150, 1996.

[13] M.R. Segal, "Features of Tree-Structured Survival Analysis," *Epidemiology*, vol. 8, (no. 4), pp. 344-346, 1997.

[14] S.S. Hwang, C.B. Scott, V.T. Chang, J. Cogswell, S. Srinivas, and B. Kasimis, "Prediction of Survival for Advanced Cancer Patients by Recursive Partitioning Analysis: Role of Karnofsky Performance Status, Quality of Life, and Symptom Distress," *Cancer Investigation*, vol. 22, (no. 5), pp. 678-687,2004.

[15] J. Bazan, A. Osmólski, A. Skowron, D. Ślęzak, M. Szczuka, and J. Wróblewski, "Rough Set Approach to the Survival Analysis - Rough Sets and Current Trends in Computing," vol. 2475, *Lecture Notes in Computer Science*, J. Alpigini, J. Peters, A. Skowron and N. Zhong eds.: Springer Berlin / Heidelberg, pp. 951-951, 2002.

[16] P. Pattaraintakorn, N. Cercone, and K. Naruedomkul, "Hybrid rough sets intelligent system architecture for survival analysis," in Transactions on rough sets VII, W. M. Victor, O. Ewa, owska, S. Roman, owinski and Z. Wojciech eds.: Springer-Verlag, 2007, pp. 206-224.

[17] F. Lau, D. Cloutier-Fisher, C. Kuziemsky, F. Black, M. Downing, and E. Borycki, *A systematic review of prognostic tools for estimating survival time in palliative care*, Montreal, CANADA: Centre of Bioethics, Clinical Research Institute of Montreal, 2007.

[18] P. Glare, C. Sinclair, M. Downing, P. Stone, M. Maltoni, and A. Vigano, "Predicting survival in patients with advanced disease," *European Journal of Cancer*, vol. 44, (no. 8), pp. 1146-1156, 2008.

[19] P. Simons, "VAGUENESS" *International Journal of Philosophical Studies*, vol. 4, (no. 2), pp. 321-327, Sep 1996.

[20] B. Djulbegovic, "Medical diagnosis and philosophy of vagueness-uncertainty due to borderline cases," *Annals of Internal Medicine*, 2008.

[21] J.C. Wyatt and D.G. Altman, "Commentary: Prognostic models: clinically useful or quickly forgotten?," *BMJ*, vol. 311, (no. 7019), pp. 1539-1541, 1995.

[22] R.L. Kravitz, N. Duan, and J. Braslow, "Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages," *Milbank Quarterly*, vol. 82, (no. 4), pp. 661-687, 2004.

[23] P. Schlattmann, "Introduction - Heterogeneity in Medicine Medical Applications of Finite Mixture Models," *Statistics for Biology and Health*: Springer Berlin Heidelberg, 2009, pp. 1-22.

[24] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Norwell, MA, 1992.

[25] S. Tsumoto, "Modelling Medical Diagnostic Rules Based on Rough Sets- Rough Sets and Current Trends in Computing," vol. 1424, *Lecture Notes in Computer Science*, L. Polkowski and A. Skowron eds.: Springer Berlin / Heidelberg, 1998, pp. 475-482.

[26] J. Komorowski and A. Øhrn, "Modelling prognostic power of cardiac tests using rough sets," *Artificial Intelligence in Medicine*, vol. 15, (no. 2), pp. 167-191, 1999.

[27] P. Grzymala-Busse, J.W. Grzymala-Busse, and Z.S. Hippe, "Melanoma prediction using data mining system LERS," in *Proc,, COMPSAC*, 2001, pp. 615-620.

[28] P. Pattaraintakorn and N. Cercone, "Integrating rough set theory and medical applications," *Applied Mathematics Letters*, vol. 21, (no. 4), pp. 400-403, 2008.

[29] E. Gil-Herrera, A. Yalcin, A. Tsalatsanis, L.E. Barnes, and D. B, "Rough Set Theory based Prognostication of Life Expectancy for Terminally Ill Patients," in *Conf Proc IEEE Eng Med Biol Soc*, 2011, pp. 6438-6441.

[30] J. Bazan, H. Nguyen, S. Nguyen, P. Synak, J. Wroblewski, L. Polkowski, S. Tsumoto, and T. Lin, "Rough Set Algorithms in Classification Problem," in Rough set methods and applications: new developments in knowledge discovery in information systems: Physica-Verlag, 2000, pp. 49-88.

[31] Ø. Alexander and J. Komorowski, "ROSETTA: A Rough Set Toolkit for Analysis of Data," in *Proc. Third International Joint Conference on Information Sciences*, 1997, pp. 403-407.

[32] J.G. Bazan, H.S. Nguyen, S.H. Nguyen, P. Synak and J. Wróblewski, "Rough set algorithms in classification problem," in *Rough set methods and applications*: Physica-Verlag GmbH, 2000, pp. 49-88.

Djulbegovic et al. BMC Medical Informatics and Decision Making 2012, 12:94
http://www.biomedcentral.com/1472-6947/12/94

BMC
Medical Informatics & Decision Making

RESEARCH ARTICLE Open Access

Dual processing model of medical decision-making

Benjamin Djulbegovic[1,2,3,7*], Iztok Hozo[4], Jason Beckstead[5], Athanasios Tsalatsanis[1,2] and Stephen G Pauker[6]

Abstract

Background: Dual processing theory of human cognition postulates that reasoning and decision-making can be described as a function of both an intuitive, experiential, affective system (system I) and/or an analytical, deliberative (system II) processing system. To date no formal descriptive model of medical decision-making based on dual processing theory has been developed. Here we postulate such a model and apply it to a common clinical situation: whether treatment should be administered to the patient who may or may not have a disease.

Methods: We developed a mathematical model in which we linked a recently proposed descriptive psychological model of cognition with the threshold model of medical decision-making and show how this approach can be used to better understand decision-making at the bedside and explain the widespread variation in treatments observed in clinical practice.

Results: We show that physician's beliefs about whether to treat at higher (lower) probability levels compared to the prescriptive therapeutic thresholds obtained via system II processing is moderated by system I and the ratio of benefit and harms as evaluated by both system I and II. Under some conditions, the system I decision maker's threshold may dramatically drop below the expected utility threshold derived by system II. This can explain the overtreatment often seen in the contemporary practice. The opposite can also occur as in the situations where empirical evidence is considered unreliable, or when cognitive processes of decision-makers are biased through recent experience: the threshold will increase relative to the normative threshold value derived via system II using expected utility threshold. This inclination for the higher diagnostic certainty may, in turn, explain undertreatment that is also documented in the current medical practice.

Conclusions: We have developed the first dual processing model of medical decision-making that has potential to enrich the current medical decision-making field, which is still to the large extent dominated by expected utility theory. The model also provides a platform for reconciling two groups of competing dual processing theories (parallel competitive with default-interventionalist theories).

Background

Dual processing theory is currently widely accepted as a dominant explanation of cognitive processes that characterizes human decision-making [1-9]. It assumes that cognitive processes are governed by so called system I (which is intuitive, automatic, fast, narrative, experiential and affect-based) and system II (which is analytical, slow, verbal, deliberative and logical) [1-10]. The vast majority of existing models of decision-making including expected-utility theory, prospect theory, and their variants assume a single system of human thought [11]. Recently, formal models for integrating system I with system II models have been developed [3,11]. One such attractive model-Dual System Model (DSM)- has been developed by Mukherjee [11]. Here, we extend Mukherjee's DSM model to medical field (DSM-M) by linking it to the threshold concept of decision-making [12-15]. We also take into account decision regret, as an exemplar of affect or emotion that is involved in system I decision-making [2], and which is of particular relevance to medical decision-making [16-19]. Regret was also selected for use in our model because any

* Correspondence: bdjulbeg@health.usf.edu
[1]Center for Evidence-based Medicine and Health Outcomes Research, Tampa, FL, USA
[2]Department of Internal Medicine, Division of Evidence-based Medicine and Health Outcomes Research University of South Florida, Tampa, FL, USA
Full list of author information is available at the end of the article

"theory of choice that completely ignores feeling such as the pain of losses and the regret of mistakes is not only descriptively unrealistic but also might lead to prescriptions that do not maximize the utility of outcomes as they are actually experienced" [1,20].

As more than 30% of medical interventions are currently not appropriately applied, mostly as over - or-undertreatment [21-23], we illustrate how the DSM-M model may be used to explain the practice patterns seen in the current medical practice. Our DSM-M model is primarily an attempt to describe how medical decisions are made. As a descriptive model its validation will require comparing its outputs to actual choices made by patients and clinicians and their verbalized reactions to our model. We conclude the paper by providing some testable empirical predictions.

Methods

A dual system model

Building on the previous empirical research, which has convincingly showed that human cognition is determined by both system I and system II processes [1,2,5,24,25]. Mukherjee recently developed a formal mathematical model, which assumes parallel functioning by both systems, while the final decision is a weighted combination of the valuations from both systems based on the value maximization paradigm (Figure 1) [11]. (NB. In this paper we employ terms system I and system II as popularized by Kahneman [1,2] although some authors prefer to talk about type 1 and 2 processing as it is almost certain that human cognition is not organized in distinctly separated physical systems [5,26,27]).

Mukherjee's dual system model (DSM) assumes that evaluation of risky choice (C) is formed by the combined input of system I and system II into a single value and can be formulated as follows:

$$E(C) = \gamma V_I(C) + (1-\gamma)V_{II}(C) = \gamma \frac{1}{n}\sum_i V_I(x_i) + (1-\gamma)k\sum_i p_i V_{II}(x_i) \quad (1)$$

Where C represents a decision-making situation ("choice"), n - number of outcomes, p_i - probability of the i^{th} outcome, x_i, of the selected choice. V_I represents valuation of decision under autonomous, intuitive, system I-based mode of decision-making and V_{II}, which can be a utility function, represents valuation under a deliberative, rule-based, system II mode of decision-making. k-is a scaling constant, and γ [0 to 1] is the weight given to system I and can be interpreted as the relative extent of involvement of system I in the decision-making process [11]. System II is not split into two subsystems advocated by some [5], but is assumed to adhere to the rationality criteria of expected utility theory (EUT) as also advocated by modern decision science [11,28]. γ is assumed to be influenced by a number of processes that determine system I functioning. Mukherjee emphasized the following factors as the important determinants of system I functioning [11]: individual decision-making and thinking predispositions [ranging from expected utility theory (EUT) "maximizers" to system I driven "satisficing" with no regard to probabilities but with editing or selection of outcomes of interest] [29], affective nature of outcomes (the higher the affective nature of outcomes, the higher is γ) and framing and construing the decision-making task (decisions for the self will likely have higher γ, as well as decision problems that are contextualized and those requiring immediate resolution or are made under time pressure; the last four describe circumstances characteristic of medical decision-making). Easily available information, our previous experience, the way in which information is processed (verbatim vs. getting the "gist" of it) [30] as well as memory limitations [31] are also expected to affect γ. γ is, therefore, expected to be higher when information about probabilities and outcomes are ambiguous or not readily available, or when a very severe negative prior outcome is recalled [2,32,33]. On the other hand, when such data are available their joint evaluation by system II will reduce γ [11]. In general, the factors that define the process of system I can be classified under 4 major categories: a) affect, b) evolutionary hard-wired processes, responsible for automatic responses to potential danger in such a way that system I typically gives higher weight to potentially false positives than to false negatives (i.e. humans are cognitively more ready to wrongly accept the signal of potential harms than one that carries the potential of benefit),

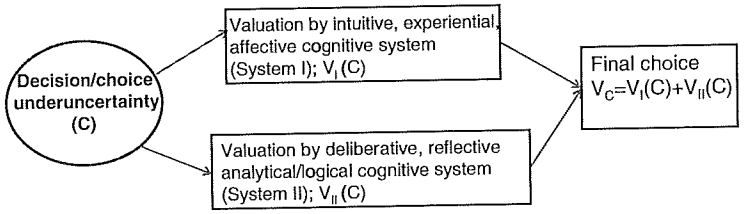


Figure 1 Model of decision-making using dual processing cognitive processes (after Mukherjee [11]).

(c) over-learned processes from system II that have been relegated to system I (such as the effect of intensive training resulting in the use of heuristics, or "rules of thumb" or practice guidelines as one of the effort-saving cognitive strategies. NB although guidelines may be the products of analytic system II processes their applications tends to be a system I process.), and (d) the effects of tacit learning [5].

Mukherjee's DSM model draws upon empirical evidence demonstrating that decision-makers in an affect-rich context are generally sensitive only to the presence or absence of stimuli, while in affect-poor contexts they rely on system II to assess the magnitude of stimuli (and probabilities) [24]. Hence, the salient feature of the model is that that system I recognizes outcomes only as being possible or, not. Every outcome that remains under consideration gets equal weight in system I. On the other hand, system II recognizes probabilities linearly without distortions, according to the expected utility paradigm.

As a result, dual valuation processing often generates instances where subjective valuations are greater at lower stimulus magnitudes (i.e. when decision-making relies on feeling, or evolutionary hard-wired processes such as when the signal may present danger) while rational calculation produces greater value at high magnitudes [11]. DSM is capable of explaining a number of the phenomena that characterize human decision-making such as a) violation of nontransparent stochastic dominance, b) fourfold pattern of risk attitude, c) ambiguity aversion, d) common consequences effect, e) common ratio effect, f) isolation effect, g) and coalescing and event-splitting effect [11].

Under the realistic assumption that outcomes are positive (i.e., utilities >0, which is particularly applicable to medical setting) and power value functions, $V_I(x) = x^{m_I}$, and $V_{II}(x) = x$ for system I and system II, respectively, DSM can be re-written as:

$$V(C) = \gamma \frac{1}{n} \sum_i x_i^{m_I} + (1-\gamma) k \sum_i p_i x_i \qquad (2)$$

where $0 < m_I \leq 1$ Note that $x_i^{m_I}$ satisfies risk aversion for gains and risk seeking for losses and that the term for system II $p_i x_i$ is linear without risk distortions.

As noted by Mukherjee [11], the estimation of the parameters in Equation 2) is a measurement exercise, which needs to be evaluated in the future empirical research. Consequently, the functions $V_{II}(x)$ *and* $V_I(x)$ could be changed, depending on the decision-making setting and decision-maker's goals. Similarly, parameter m may not be the same for all outcomes.

Modification of DSM for medical decision-making

We will consider a typical situation in clinical decision-making where a doctor has to choose treatment (Rx) vs. no treatment (NoRx) for disease (D) which is present with the probability p. [Note than NoRx represents a competing treatment alternative and may include a different treatment (Rx2)] [12,34]. Each decision results in outcomes that have a certain value, x_i. The model is shown in the Figure 2. As noted above, the system I recognizes outcomes only as being possible (or not), and is thus insensitive to exact probabilities. Every outcome with non-zero probability gets equal weight in system I. Hence, in a two-alternative choice, each probability is equal to 0.5 under system I. System II recognizes probabilities without distortions, as would be expected according to EUT.

We posit that among the emotions that can influence valuation of outcomes in system I processing, regret plays

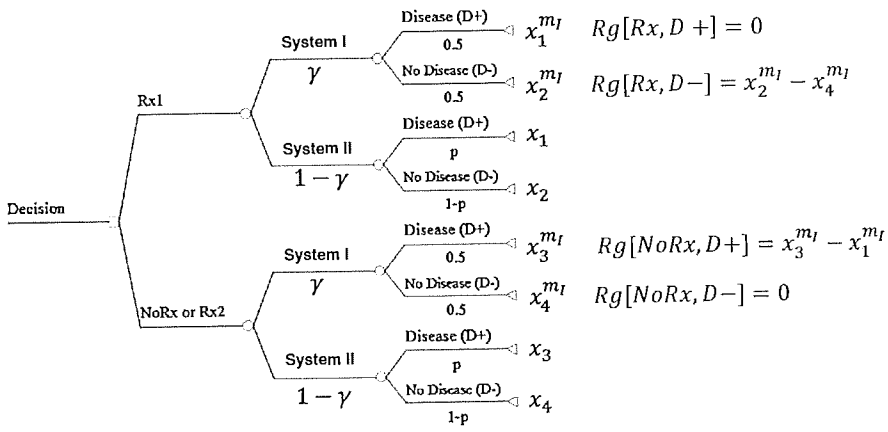


Figure 2 Dual processing model of decision-making as applied to a clinical dilemma whether to treat (Rx) the patient with disease (D+) or not. The patient may or may not have a disease (probability *p*). Regret is assumed to operate at the level of system I only. (Note that competing treatment alternative may include Rx or NoRx). Rg- regret.

an important role [1,2], while system II processes are dominated by rational, analytical deliberations according to EUT [11]. We can define regret (Rg) as the difference (loss) in the utilities of the outcome of the action taken and that of the action we should have taken, in retrospect [16-19,35] but operating at the system I level only (see Figure 2).

Hence, we have the following value functions (see Additional file 1: Appendix for detailed derivation):

$$\begin{aligned} V_I(Rx, D+) &= Rg[Rx, D+] = 0; \\ V_I(NoRx, D+) &= Rg[NoRx, D+] = x_3^{m_I} - x_1^{m_I}; \\ V_{II}(Rx,\ D+) &= x_1\ ; \\ V_{II}(NoRx,\ D+) &= x_3; \end{aligned}$$

$$\begin{aligned} V_I(Rx, D-) &= Rg[Rx, D-] = x_2^{m_I} - x_4^{m_I} \\ V_I(NoRx,\ D-) &= Rg[NoRx,\ D-] = 0; \\ V_{II}(Rx,\ D-) &= x_2; \\ V_{II}(NoRx,\ D-) &= x_4; \end{aligned}$$

Overall valuation of decision to treat (Rx) is equal to:

$$\begin{aligned} V(Rx) &= \frac{\gamma}{2}(V_A(Rx, D+) + V_A(Rx,\ D-)) \\ &\quad +(1-\gamma)k(pV_D(Rx, D+) \\ &\quad +(1-p)V_D(Rx,\ D-)) \\ &= \frac{\gamma}{2}(x_2^{m_A} - x_4^{m_A}) + (1-\gamma)k[px_1 + (1-p)x_2] \end{aligned}$$

And

$$\begin{aligned} V(NoRx) &= \frac{\gamma}{2}(V_A(NoRx, D+) + V_A(NoRx,\ D-)) \\ &\quad +(1-\gamma)k(pV_D(NoRx, D+) \\ &\quad +(1-p)V_D(NoRx,\ D-)) \\ &= \frac{\gamma}{2}(x_3^{m_A} - x_1^{m_A}) + (1-\gamma)k[px_3\ +(1-p)x_4] \end{aligned}$$

The difference in the outcomes of treating and not treating patient with disease are equal to the net benefit of treatment (B) [13,14,36]; the difference in outcomes of not treating and treating those patients without disease is defined as net harms (H) [13,14,36]. Note that benefits and harms can be expressed in the various units (such as survival, mortality, morbidity, costs, etc.) and can be formulated both as utilities and disutilities [13,14,36]. As explained above, we further assume that valuation of net benefits and net harms by system I differs from system II. Hence, under system II, we replace net benefit and net harms using EUT definitions:$B_{II} = x_1 - x_3$ and net harms $H_{II} = x_4 - x_2$. Under system I, we define $B_I = x_1^{m_I} - x_3^{m_I}$, and $H_I = x_4^{m_I} - x_2^{m_I}$. Solving for p (the probability of disease at which we are indifferent between Rx and NoRx), we obtain: (Equation 3)

$$\begin{aligned} p_t = p &= \frac{(1-\gamma)kH_{II} - \frac{\gamma}{2}[B_I - H_I]}{(1-\gamma)k[B_{II} + H_{II}]} \\ &= \frac{1}{1+\frac{B_{II}}{H_{II}}} - \frac{\gamma}{2k(1-\gamma)}\frac{B_I - H_I}{B_{II} + H_{II}} \\ &= \left(\frac{1}{1+\frac{B_{II}}{H_{II}}}\right)\left[1 + \frac{\gamma}{2(1-\gamma)}\left(\frac{H_I}{H_{II}}\right)\left(1 - \frac{B_I}{H_I}\right)\right] \\ &= (p_t(EUT))\left[1 + \frac{\gamma}{2(1-\gamma)}\left(\frac{H_I}{H_{II}}\right)\left(1 - \frac{B_I}{H_I}\right)\right] \end{aligned} \quad (3)$$

This means that if the probability of disease is above p_t the decision-maker favors treatment; otherwise, a competing management alternative (such as "No Treatment") represents the optimal treatment strategy. Note that k can be typically set at 1, as we do it here. Also note that the first part of equation is equivalent to the threshold expression described in EUT framework [13,14,36]; the second expression modifies system II's EUT-based decision-making process in such a way that if benefits are experienced higher than harms, the threshold probability is always lower than EUT threshold. However, if a decision-maker experiences $H_I > B_I$, the threshold probability is always higher than the EUT threshold (see below for discussion in the context of medical example). Note that γ and the ratio $\frac{H_I}{H_{II}}$ only contribute to the extent of magnitude the dual threshold is above or below the classic EUT threshold. That is, γ and the ratio $\frac{H_I}{H_{II}}$ do not change the quality of relationship between dual threshold and EUT threshold: whether dual threshold will be above or below the EUT threshold depends only on a $\frac{B_I}{H_I}$ ratio.

It should be noted that the identical derivations can be obtained by applying the concept of expected regret (instead of EUT) [16-19,35]. Although it can be argued that regret is a powerful emotion influencing all cognitive processes (as so called, "cognitive emotion") [37,38], and so it may function at level of both system I and system II [39], most authors recognize the affect value of regret [2,10]. Hence, we assumed that regret functions at system I level [2]. Therefore, in our model we restrict the influence of regret to system I. Incidentally, our Equation 3) can also be derived from the general Mukherjee's DSM model even if regret is not specifically invoked [11].

Although Equation 3) implies exact calculations, it should not be understood as one that provides precise mathematical account of human decision-making. Rather, it should be considered more as a semi-quantitative or qualitative description of the way physicians may make their decisions. First, this is because system I does not perform exact calculations, but rather relies on "gist" [30,31]

for assessment of benefits and harms in more qualitative manner. The mechanism depends on associations, emotions (so called, "risk as feelings" estimates [10]), as well as memory, and experience [2,5,8,31]. In this sense, the second part of Equation 3) that relies on system I can be understood as the qualitative modifier ("weight"), which, depending on the system I's estimates of benefits and harms increases or decreases the first part of equation (which is dependent on system's II precise usage of evidence for benefits and harms). Second, the threshold probability itself should be considered as an "action threshold"- at some point, a physician decides whether to administer treatment or not. Typically, she contrasts the estimated probability of disease against the threshold and acts: if the probability of disease is above the "action threshold", the physician administers the treatment; if it is below, she decides not to give treatment. So, one way to interpret Equation 3) is to consider physician's estimate of "gist" of the action threshold: if in her estimation, overall benefits of treatment outweigh harms, and she considers that it is "likely" that the probability of disease is above the threshold probability, then she would act and administer treatment. If the physician assesses that it is "unlikely" that the probability disease is above the "action threshold", then she would not prescribe the treatment.

The behavior of DSM-M model

The exact cognitive mechanisms that underlie dual system processes are not fully elucidated. As discussed throughout this paper, many factors affect dual processes reasoning leading to suggestions that these processes should be grouped according to the prevailing mechanisms [27]. Focusing on each of these processes may lead to specific theoretical proposals. Our goal in this paper is to provide overarching cognitive architecture encompassing general features of the majority existing theoretical concepts, while at the same time concentrating on specifics of medical decision-making. In general, dual processing theories [27] fall into two main groups [27,40] parallel competitive theories and default-interventionalist theories. The parallel-competitive theories assume that system I and II processes proceed in parallel, each competing for control of the response [27]. If there is a conflict, it is not clear which mechanism is invoked to resolve the conflict [27]. On the other hand, default-interventionist theories postulate that system I generates a rapid and intuitive default response, which may or may not be intervened upon by subsequent slow and deliberative processed of system II [2,5,27]. This can be further operationalized via several general mechanisms that have been proposed in the literature:

1) Mukherjee's additive model as described above [11]. It can be categorized as a variant of parallel-competitive theory as it assumes that system I and II processes proceed in parallel, but does include parameter γ, which can trigger greater or smaller activation of system I. Mukherjee's model, however, does not explicitly model the choices in terms of categorical decisions (i.e. accept vs. do not accept a given hypothesis), which is a fundamental feature of dual-processing models [27].
2) System I and system II operate on a continuum [41], but in such a way that system I never sleeps [2]. A final decision depends on the activation of both systems I and II [2]. It has been estimated that about 40-50% of decisions are determined by habits (i.e. by system I) [42]. This is also a variation of parallel-competitive theory; it should be noted that latest literature is moving away from this model [5,27].
3) The final decision appears to depend both on the system I and system II in such a way that system I is the first to suggest an answer and system II endorses it [2]. In doing so, system II can exert the full control over system I (such as when it relies on the EUT modeling) or completely fail to oversee functioning of system I (e.g., because of its ignorance or laziness) [2]. Therefore, according to this model, decisions are either made by system I (default) or system II (which may or may not intervene). This is a default-interventionalist model.
4) The variation of the model #3 is the so called "toggle model", which proposes that decision-maker constantly uses cognitive processes that oscillate between the two systems (toggle) [6,7,9]. This is a variant of default-interventionalist model.

Note that γ is continuous in our model, but it can be made categorical [0,1] if the "toggle" theory is considered to be the correct one. In this case, a logical switch can be introduced in the decision tree to allow toggling between the two systems. Most importantly, by linking Mukherjee's additive model with the threshold model, we provide the architecture for reconciling parallel competitive theories with default-interventionalist theories. We do it by making explicit that decisions are categorical (via threshold) at certain degree of cognitive effort (modeled via γ) parameter [27]. That is, the key question is what processes determine acceptance or rejection of a particular (diagnostic) hypothesis. Our model shows that this can occur if we maintain parallel-competing architecture of Mukherjee's additive model but assume a switch, yes or no answer, whether to accept or reject a given hypothesis (*at the threshold*). It is evaluation of the (diagnostic) event with respect to the threshold that serves as the final output of our decision-making and reasoning processes. As our model shows, this depends on assumption of parallel working of both system I and system II, *and* the switch in control of one system over another according to default-

interventionalist hypothesis. Note that depending on activation of γ parameter and assessment of benefits (gains) and harms (losses) the control can be exerted by either system: sometimes it will be the intuitive system that it will exert the control and our action will take the form "feeling of rightness" [43]; sometimes, it will be system II that it will prevail and drive our decisions. Thus, we succeed in uniting parallel competitive with default-interventionalist models by linking Mukherjee's additive model with the threshold model for decision-making.

As discussed above, many factors can activate the switch such as the presence or absence of empirical, quantitative data, the context of decision making (e.g. affect poor or rich), the decision maker's expertise and experience, etc. In addition, extensive psychological research has demonstrated that people often use a simple heuristic, which is based on the prominent numbers as powers of 10 (e.g., 1,2,5,10,20,50,100,200 etc.) [44]. That is, although system I does not perform the exact calculations, it still does assess "gist" of relative benefits and harms, and likely does so according to "1/10 aspiration level" [44] (rounded to the closest number) in such a way that the estimates of benefits/harms ratio change by 1,2,5, 10, etc. orders of magnitude. Therefore, in this section we consider several prototypical situations: 1) when γ = 0, 0.5, or 1; 2) when $B_{II} >> H_{II}$, $B_{II} = H_{II}$ and $B_{II} << H_{II}$; and 3) when regret of omission (B_I) << regret of commission (H_I), $B_I = H_I$, or $B_I >> H_I$

First, note that γ=0, when the numerator of the left fraction in the Equation 6 (Additional file 1: Appendix) is zero, i.e., when $pB_{II} - (1-p)H_{II} = 0$, or solving for p, we obtain $p = \frac{1}{1+\frac{B_{II}}{H_{II}}}$, which is exactly the value of the EUT threshold for the probability at which the expected utilities of the two options are the same. This will correspond to model #3 above, in which system II exerts full control over decision-making. Therefore, when γ = 0, we have the classic EUT and therapeutic threshold model. In this case, regret does not affect the EUT benefits and harms, and $p_t = \frac{H_{II}}{H_{II}+B_{II}} = \frac{1}{1+\frac{B_{II}}{H_{II}}}$. If $B_{II} >> H_{II}$, p_t approaches zero and a decision-maker will recommend treatment to virtually everyone. On the other hand, if $B_{II} = H_{II}$, p_t equals 0.5 and she might recommend treatment if the disease is as likely as not. Finally, if $B_{II} << H_{II}$, p_t approaches 1.0, and the decision-maker is expected to recommend treatment only if she is absolutely certain in diagnosis.

At the other extreme, if γ = 1, we have the pure system I model (corresponding to model #3 above, which solely relies on system I processes). Note the value of γ=1, when the denominator of the second fraction in Equation 6 (Additional file 1: Appendix) equals one, or when the expression $H_I - B_I = 0$, i.e., when B_I=H_I. Under these conditions, it is fairly obvious that the system I assessments become irrelevant if the perceived net benefit of the treatment is equal to the perceived net harm. When γ=1, regret avoidance becomes the key motivator, not EUT's benefits and harms. Note that in system I p is not related to γ in terms of the valuation (Equation 1). Under these circumstances only decision-making under system I operate and the analytical processes of system II are suppressed (Equation 1) as seen in those decision-makers who tend to follow intuition only, or are extremely affected by their past experiences without considering new facts on the ground. That is, differences in probability do not play any role in such decisions, because a person who only uses system I doesn't consider probability as a factor.

Finally, if γ = 0.5, the decision maker is motivated by EUT and by regret avoidance (model #2 listed above). In this case, the benefits (B_{II}), harms (H_{II}), regrets of omission (B_I) and commission (H_I) are all active players. These three cases are presented in Table 1 (see Additional file 2) which shows threshold probabilities for γ = 0.5 and objective data indicating a high benefit/harms ratio ($B_{II}/H_{II} = 10$). Also shown is how the threshold probability depends on individual risk perception. If $H_I >> H_I$, it magnifies effect of B_I/H_I (see Equation 3), which results in extreme behavior in sense of increasing likelihood that such a person will either always accept (as p_t<0) or reject treatment (as p_t>1). For $H_I << H_{II}$, the impact on the way system I processes benefits and harms is not that pronounced and influences the EUT threshold to much smaller extent.

Results

Illustrative medical examples

Clinical examples abound to illustrate applicability of our model. To illustrate the salient points of our model, we chose two prototypical examples where there is close trade-offs between treatments' benefits and harms.

Example #1: treatment of pulmonary embolism

Pulmonary embolism (PE) (blood clot in the lungs) is an important clinical problem that can lead to significant morbidity and death [45]. Even though many diagnostic imaging tests exist to aid in the accurate diagnosis of PE, the tests are often inconclusive, and physicians are left to face the decision whether to treat patient for presumptive PE, or attribute the patient's clinical presentation (such as shortness of breath and/or chest pain) to other possible etiologies. There exists an effective treatment for a PE, which consists of the administration of 2 anticoagulants (blood thinners): heparin followed by oral anticoagulants such as warfarin [46,47]. Heparin (unfractionated or low-molecular weight heparins) are highly effective treatments associated with relative risk reduction of death from PE by 70-90% in comparison to no treatment [46,47]. This converts into the absolute death reduction as: net benefits,

B_{II}=17.5% *to* 22.5% (calculated as 25% morality without heparin minus 7.5% to 2.5% with heparin) [17,18,46,47]. However, these drugs are also associated with a significant risk of life-threatening bleeding; net harms range from H_{II}=0.037% (a typical scenario) to 5% (a worst-case scenario) depending on the patients' other comorbid conditions [17,18,47,48]. Thus, net benefits/net harms range from 60.8 (22.5/0.037) (best case) to 3.5 (17.5/5)(worst case scenario). If we apply a classic EUT threshold [13,14,36], which relies solely on system II processes, we observe that the probability of pulmonary embolism above which the physician should administer anticoagulants ranges from 1.6% $[=1/(1+60.8)]$ (best case) to 22.2% $[1/(1+3.5)]$(worst case scenario). However, ample clinical experience has demonstrated that few clinicians would consider prescribing anticoagulants at such low probability of PE [18]. In fact, most experts in the field recommend giving anticoagulants when probability of PE exceeds 95% [49-51]. We have previously suggested that this is because regret associated with administering unnecessary and potentially harmful treatments under these circumstances likely outweighs regret associated with failing to administer potentially beneficial anticoagulants [17-19]. We now show how this argument can be made in the context of dual processing theory. Indeed, some physicians may feel that the risk of bleeding may be much higher, particularly in case of a patient who recently experienced major hemorrhage. The physician may not have data readily available to adjust her EUT, system II-based calculations. Rather, she employs the system I-based reasoning, globally assessing the benefits and harms of treatments under her disposal. Importantly, these are personal, intuitive, affect-based, subjective judgments of the values of outcomes that are influenced by memory limitations and recent experiences and that may not be objectively based on the external evidence [2,30-33]. In addition, it is well documented that the physicians' recent experience leads to a type of bias, known as primacy effect, that is governed by system I [2,33]. If the last patient with PE whom the physician took care of had severe bleeding, system I may be primed in such a way that it will likely conclude that harms outweigh benefits. In our case of PE, if her reasoning is dominated by system I (operating, say, at γ level of 0.77 according to model #2 listed above, see Section "The behavior of DSM-M model") in a such way that the physician concludes that if harms is larger than benefits by 10%, then the threshold probability above which she will treat her patient suspected of PE exceeds 95% [as easily demonstrated after plugging in the benefits/harms values in Equation 3) $p_t(\text{dual}) = .222 - (.77)/(2^*.23)^*(-.10/.225) = 0.966 =$ 96.6% for $k = 1$]. Note that this calculation describes circumstances under which the physician would adhere to the contemporary practice guidelines i.e. to prescribe anticoagulants when PE exceeds 95% [49-51]. It should be further noted that if γ value is only slightly higher (≥0.78), the physician will require the absolute certainty to act (i.e. the threshold ≥1).

DSM offers an account of the opposite behavior as well i.e. the threshold based on global evaluation using both system I and system II can also be lower than the EUT threshold (if $B_I > H_I$ additive, model #1, Equation 3). For example, the physician may trivialize the risks of treatment and believe that the benefits are much higher than the treatment harms. As a result, the threshold above which she commits to treatment drops below EUT threshold (as predicted by Equation 3). Figure 3 shows how the decision threshold (p_t) is affected by the relative involvement of systems I and II in dual process model of medical decision-making in the "best" ($B_{II}/H_{II} = 60.8$) and "worst case" scenario ($B_{II}/H_{II} = 3.5$) for treatment of PE and when system I valuation of benefits is greater than harms or when harms are perceived to outweigh benefits. It can be seen that when objective data indicate that benefits considerably outweigh harms ($B_{II} >>> H_{II}$) (as when $B_{II}/H_{II} = 60.8$), then as long as system I values benefits as being greater than harms, the threshold dramatically drops to zero indicating that the extent of system I involvement (i.e. γ value) in decision-making is of little consequence. However, if system I clashes with objective data, then the probability of PE above which the decision-maker is prepared to treat, dramatically increases (Figure 3a). Similarly, in all other circumstances (when $B_{II} > H_{II}$, $B_{II} \sim H_{II}$, $B_{II} < H_{II}$), the threshold probability is significantly affected by involvement of system I (Figures 3b–3d).

Example #2: treatment of acute leukemia

Acute myeloid leukemia (AML) is a life-threatening disease, which, depending on the aggressiveness of disease can be cured in the substantial minority of patients. To achieve a cure, patients are typically given induction chemotherapy to bring the disease into remission, after which another form of intensive therapy - so called, consolidation treatment - is given. To achieve a cure in patients with more aggressive course of disease such as those classified as intermediate- and poor-risk AML based on cytogenetic features of disease, allogeneic stem cell transplant (alloSCT) is recommended [52]. However, the cure is not without price- many patients given alloSCT as a consolidation therapy die due to treatment. A decision dilemma faced by a physician is whether to recommend alloSCT, or alternative treatment, such as chemotherapy or autologous SCT, which has lower cure rate but less treatment-related mortality. In intermediate-risk AML, for example, credible evidence shows that, compared with chemotherapy allogeneic alloSCT result in better leukemia-free survival (LFS) by at least 12% at 4 years (LFS with alloSCT =53% vs 41% with chemotherapy/auto SCT) [53]. Treatment-related

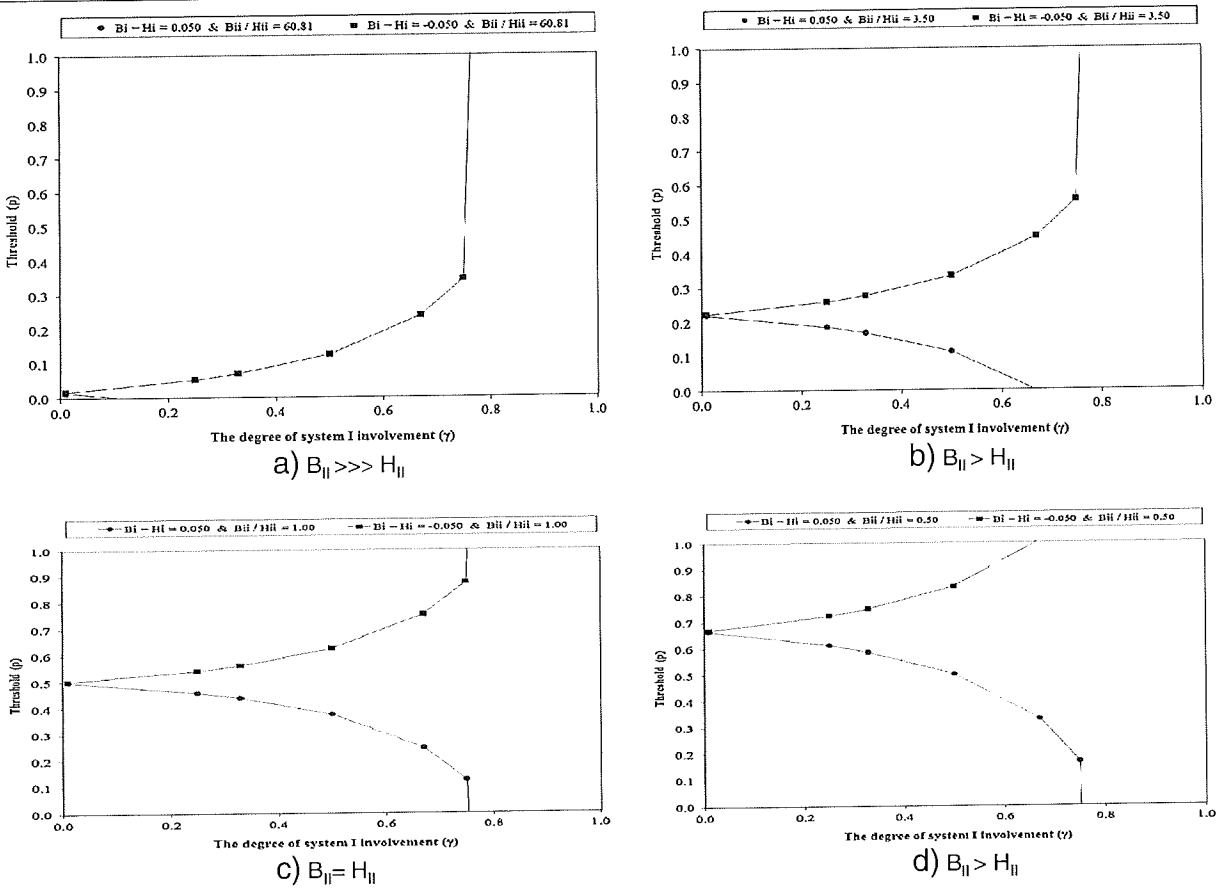


Figure 3 Dual decision threshold (p_t) as the function of relative involvement of systems I and II in dual process model of medical decision making: a) objective data show very high benefit/harms ratio (B_{II}> > H_{II}), b) moderately high benefit/harms ratio (B_{II} > H_{II}), c) B_{II} = H_{II}, d) B_{II} < H_{II}. The intercept at y axis threshold probability. The graph shows how the threshold is affected by the extent of system I involvement (γ) and whether system I perceives that benefits is greater than harms [by 5% in this example](red lines, circles) or that harms outweigh benefits[by 5%] (blue line, squares). Decision-maker accepts treatment if the probability of disease exceeds the threshold; otherwise, treatment would not be acceptable (see text for details).

mortality is much higher with alloSCT by 16%, on average (19% with alloSCT vs. 3% with chemotherapy/autoSCT) [53]. This means that based on objective data, and using rational EUT model, we should recommend alloSCT for any probability of AML relapse ≥57.1% (threshold = $1/(1 + 0.12/0.16) = 0.571$). Therefore, treatment benefits and harms are, on average, very close. Because of this, the driving force to recommend alloSCT is the physician's estimates of the patient's tolerability of alloSCT: if she assess that the patient will not be able to tolerate alloSCT, the physician will not recommend transplant. Conversely, if she thinks that the patient will be able to tolerate allo SCT, the physician will recommend it. Although there are objective criteria to evaluate a patient's eligibility for transplant, the assessment to the large extent depends on physicians' judgment and experience [54]. That is, the assessment of patient's eligibility for transplant depends both on the objective data on benefits and harms (system II ingredients) and intuitive, gist type of judgment (characteristics of system I). As discussed above, system I does not conduct the precise calculations. Rather, it relies on "gist" or on simple heuristics such as those that are based on powers of 10 (e.g., 1,2,5,10,20, etc.) [42]. The physician, therefore, adjusts the threshold above or below based on her intuitive calculations. For instance, it is often the case that the physician whose patient recently died during the transplant is more reluctant to recommend the procedure even to those patients who, otherwise, seems fit for it. In doing so, the physician in fact modifies her/his dual system threshold upwards. In our example, let's assume that the physician judges that the harms of alloSCT for a given patient is twice as large as reported in the studies where patients were carefully selected for transplant [52]. That, in our case, would mean that mortality due to alloSCT is 32% (instead of 16%). We can

now plug these numbers in Equation 3) ($B_{II} = 0.12$, $H_{II} = 0.16$, $B_I = 0.12$, $H_I = 0.32$).

Note that the physician can make this judgment at various level of activation of system I. If the decision is predominantly driven by system I judgment then our physician's threshold according to Equation 3) is greater than 100% for all circumstances in which γ value exceeds 55%. That means that under these circumstances of system I activation, the physician will never recommend transplant. The opposite can occur for those physicians whose experience is not affected by poor patients' outcomes. Under such circumstances, the physician may judge the patient to be in such a good condition that she may re-adjust the reported treatment-related transplant risk to be as half of those observed risks in the published clinical studies (i.e. 8%). The new numbers required to determine the threshold according to Equation 3 are: $B_{II} = 0.12$, $H_{II} = 0.16$, $B_I = 0.12$, $H_I = 0.08$. If the physician relies excessively on system I, as often seen in busy clinics where decisions are routinely made on "automatic pilot", the dual threshold drops to zero (for all $\gamma > 89\%$). That means, that the physician will recommend alloSCT to all her/his patients under these circumstances.

As discussed above, we provide the precise calculations only to illustrate the logic of decision-making. The process should be understood more along semi-quantitative or qualitative description of clinical decision-making. Although currently the Equation 3) allows entry of almost any value for benefit and harms, it is probably the case that benefit and harms as perceived by system I are based on "1/10 aspirational level" [44], so that only values of 1,2,5,10, 20 etc. should be allowed. This is, however, empirical question that should be answered in further experimental testing; therefore, at this time, we decided not to provide the exact boundaries of the values for benefit and harms that can be entered in Equation 3 (see Discussion). Note also that these calculations are decision-maker specific, and although we illustrate them from the perspective of the physician, the same approach applies to the patient, who ultimately has to agree –based on her own dual cognitive processing- on the suggested course of treatment actions.

Discussion

Models of medical decision-making belong to two general classes-descriptive and prescriptive. The former, which the DSM-M exemplifies, attempt to explain why decision makers take or might take certain actions when presented with challenging decision problems abundant in contemporary medicine. The latter, exemplified by the normative therapeutic threshold models [13,14] prescribe the choice options that a rational decision maker should take. We have defined the first formal dual-process theory of medical decision-making by taking into consideration the deliberative and the experiential aspects that encompass many of the critical decisions physicians face in practice. Mathematically, our model represents an extension of Mukharjee's additive Dual System Model [11] to the clinical situation where a physician faces frequent dilemmas: whether to treat the patient who may or may not have the disease, or choose one treatment over another for prevention of disease that is yet to occur. Our model is unique in that incorporates an exemplar of strong emotion, decision regret, as one of the important components of system I functioning. We focused on regret because previous research has shown that people often violate EUT prescribed choice options in an effort to minimize anticipated regret [1,2,20]. Although we use the more common psychological term "regret," the concept is analogous to Feinstein's term "chagrin" [55]. In fact, explicit consideration of post-choice regret in decision making has been considered an essential element in any serious theory of choice and certainly dominates many clinical decisions [1,2,20]. We also reformulated the original model using the threshold concept- a fundamental approach in medical decision-making [13,14,36]. The threshold concept represents a linchpin between evidence (which presents on the continuum of credibility) and decision-making, which is a categorical exercise (as choice options are either selected or not) [13,14,36]. Using an example such as pulmonary embolism, we have shown how the extended model can explain deviations from outcomes predicted by EUT, and account for the variation in management of pulmonary embolism [45]. In general, it is possible that the huge practice variation well documented in contemporary medicine [56-61], can be, in part, due to individual differences in subjective judgments of disease prevalence and "thresholds" at which physicians act. [17,18,62]. This may be because quantitative interpretations of qualitative descriptors such as rarely, unlikely, possible, or likely [63] differ markedly among individuals and hence "gist" representations of a given clinical situation can vary widely among different physicians [30]. We are, of course, aware that many other factors contribute to variation in patient care including the structure of local care organizations, the availability of medical technologies, financial incentives etc [60]. Our intent in this article is to highlight, yet another important factor- individual differences in risk assessment as shaped by different mechanisms operating within a dual process model of human cognitive functioning [5].

There are many theories of decision-making [64]. Most assume a single system of human reasoning [11]. Nevertheless, all major theories of choice agree that rational decision-making requires integrations of benefits (gains) and harms (losses). EUT vs. non-EUT theories of decision-making differ in how benefits and harms should be integrated in a given decision task. To date, dual

processing theory provides the most compelling explanation how both intuitive and rational cognitive processes integrate information on benefits and harms and provide not only descriptive assessments of decision-making, but possibly may lead to insights that improve the way decisions are made. Figures 3 & 4 illustrate how dual decision threshold (shown on the Y axis) for deciding between two possible courses of action can be influenced by the degree of system I involvement. As discussed above and mathematically captured in Equation 3, the clinical action such as treat versus no treat is best explained by relating benefit and harms of proposed therapeutic interventions to the threshold probability: if the estimated probability of disease is greater than the threshold probability, then the decision-maker is inclined to give treatment; if the probability of disease is below the threshold, then the treatment is withheld. Figure 4 shows a dramatic drop in the decision threshold as a function of the ratio between benefit and harms, which is derived from empirically obtained evidence. When these data are solely relied on by system II, the rational course of action consists of administering treatment as long as the probability of disease is above the threshold regardless how low the threshold probability drops [13,14,36] (which in case of the treatment of a patient with pulmonary embolism can be as low as 1.6%!). Paradoxically, if we were to adopt this – presumably most rational-approach to the practice of medicine, we would likely see a further explosion of inappropriate and wasteful use of health care resources [18,21]. This is because in today's practice, benefits of approved treatments vastly outweigh their harms, and as a result threshold probability values is predictably very low for the majority of health care interventions employed in the contemporary clinical practice [18]. System I, however, does offer a means of mitigation. The correction of the thresholds - our action whether we are comfortable treating at higher or lower probability than the thresholds obtained via usage of system II – depends on the extent of involvement of system I in decision-making. If system I perceives that harms are higher than system I benefits, the threshold probability is always higher than classic EUT threshold. However, if $B_I > H_I$, the threshold probability is always lower than the EUT threshold (Figure 4). This is particularly evident in clinical practice when physicians attempt to tailor evidence based on the results of the research study, which generates the "group averages", to individual patients who often differ in important ways from patients enrolled in the research studies (e.g., these patients may be older, have comorbid conditions, might be using multiple medications, etc.) [65]. It is under these circumstances that system I affects our judgments and can give rise to different decisions from those based solely on system II. Note, however, that although system I does assess benefits and harms, it likely does so via"gist" representation and not necessarily by employing the exact numerical values as system II does [30]. System I is also affected by emotions, as illustrated in the case where experts panels of the governments of many countries recommended H1N1 influenza vaccination, but where inoculation was refused by the majority of patients [66,67].

It is interesting to examine circumstances under which we always treat ($p_t \leq 0$) or never treats ($p_t \geq 1$). Equation 1 (Additional file 2: Table S1) shows that when objective evidence indicates that benefits outweigh harms, and when this is further augmented by the decision-maker's risk attitude in such a way that it magnifies system I's valuation of benefits and harms, then we can expect to continue to witness further overtreatment in clinical practice (as p_t drops to zero) [65]. However, when the decision-maker perceives the benefits smaller than

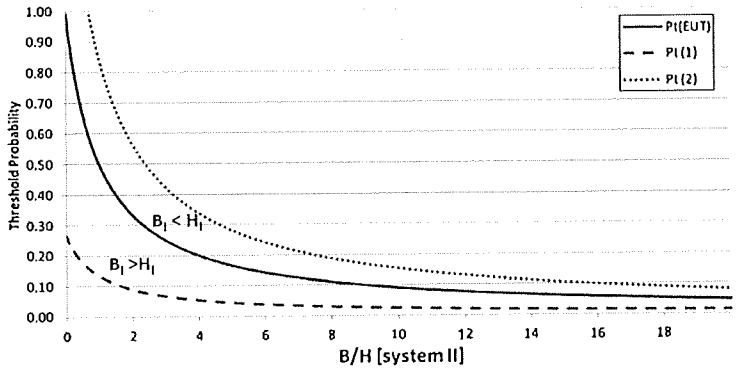


Figure 4 Dual Decision Threshold Model. Classic, expected utility threshold probability as a function of benefit/harms ratio as derived by system II, EUT (expected utility threshold) (solid line). The treatment should be given if the probability of disease is above the threshold, otherwise should be withheld. Note that if system I perceives that harms are higher than system I benefits ($B_I < H_I$), the threshold probability is always higher than classic EUT (dotted line). However, if $B_I > H_I$, the threshold probability is always lower than the EUT threshold (dashed line) (see text for details).

harms, then the threshold increases; consequently, the decision-maker will require higher diagnostic certainty before acting (Figure 3 & Figure 4). This may occur during extrapolation of research results from the group averages to individual patients, when empirical evidence about B_{II} and H_{II} is considered to be unreliable, when the decision-maker is risk averse, or when his or her cognitive processes are biased through the distorting effects of recent experience, memory limitations or other forms of biases well described in the literature [2,31,33]. This discussion illustrates how the "rationality of action" may require a re-definition, one encompassing both the formal principles of probability theory and human intuitions about good decisions [5,68]. Our goal here is not to demonstrate that one approach is conclusively superior to the other- we are merely outlining the differences in the current physicians' behavior from the perspective of dual processing theory.

Despite the growing recognition of the importance of dual processing for decision-making [2,5], a few formal models have been developed to try to capture the essence of the way we make decisions. Because different authors focus on different aspects of a multitude of decision-making processes, Evans has recently pointed out that there are many dual processing theories [27] which fall into two main groups [27,40] parallel competitive theories and default-interventionalist theories. While the exact accounts of cognitive processes between these two groups of theories differ [27], as discussed above *(Section The behavior of DSM-M Model)*, we, for the first, time provide a platform, albeit the theoretical one, for reconciling parallel competitive theories with default-interventionalist theories.

Nevertheless, our main goal is to define a theoretical model for medical decision-making; such a model may enable creation of new theoretical frameworks for future empirical research. Future research, obviously, involves extension of the model described herein to more complex clinical situations beyond relatively simple two-alternative situation, even if the latter is frequently encountered in practice. Particularly interesting will be the extension of our dual processing model to include the use of diagnostic tests as the number of new diagnostic technologies continues to explode. Finally, and most importantly, the model presented here needs empirical verification. This limitation is not unique to our model, however, and this criticism can be leveled against most current medical decision-making models, which are rarely, if ever, subjected to empirical verification.

Our model heavily relies on Mukherjee's model [11], and is accurate to the extent his additive dual processing model is correct (Figure 1, Equations 1 & 2). Also, note that we have extended Mukherjee's DSM model by omitting his scaling constant k and using general utility expressions, rather than a single parameter monotonic power function. As discussed above, many factors can activate the switch of system II. In fact, Kahneman warns [2] that "because you have little direct knowledge what goes on in your mind, you will never know that you might have made a different judgment or reached a different decision under very slightly different circumstances". This implies that the multiple factors affecting the gamma parameter cannot be directly modeled. A possible solution –and area for future research building on the psychological "fuzzy trace theory" [30]-would be to employ a fuzzy logic model to assess the values of γ (and threshold) as a function of multiple fuzzy inputs [69].

The complexity described here notwithstanding, we believe that the empirical verification of our current dual processing model is feasible. Even without direct modeling of all factors affecting γ parameter, our model generates empirically falsifiable qualitative predictions as it clearly identifies circumstances under which the *decision threshold* is increased or decreased as a function of activation of system I (γ parameter). Using simulation to imitate the various real-life decision-making scenarios [70] offers most logical avenue toward the first empirical testing of our model.

Our model also holds promise in medical education. As highlighted in Introduction, modern knowledge of cognition has taught us that most people, including physicians process information using both system I (fast, intuitive) and system II (slow, deliberative) reasoning at different times but few investigators have examined how to teach physicians to integrate both modes of reasoning in arriving at therapeutic strategies. On the diagnostic side, many investigators [6,71] have examined clinical reasoning and proposed how experienced physicians move between system I and system II, although most early papers used different terminology. The integration of system I and system II in therapeutic decision making in medicine has been less well examined. A number of investigators have proposed approaches to using and teaching system II reasoning, including the use of decision models [71]. Although this is taught in some schools it has not yet taken medical education by storm [71]. In the field of economic analysis Mukerjee has proposed a theoretical means of combining system I and system II reasoning. In this paper, we build on Mukurjee's work and show how the integration of system I and system II therapeutic reasoning can form a basis for teaching students and experienced physicians to recognize and integrate system I and system II reasoning. Our model uniquely captures most salient features of (medical) decision-making, which can be effectively employed for didactic purposes. It is believed that by recognizing separate roles of system II and the influence of system I mechanisms on the way we make decisions, we can be in a better

position to harness both types of processes toward better practice of making clinical decisions [2,9].

Conclusion

We hope that our model will stimulate new lines of empirical and theoretical work in medical decision-making. In summary, we have described the first dual processing model of medical decision-making, which has potential to enrich the current medical decision-making field dominated by expected utility theory.

Additional files

Additional file 1: Appendix: Derivation of DSM-M equation.

Additional file 2: Table S1. Evaluation of Behavior of Dual Processing Model for Medical Decision-Making (DSM -M). Threshold probability as a function of individual risk perception.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BD had an idea for the study. BD & IH jointly developed the model. IH solved the model. BD and IH performed the analyses. JB and AT performed additional analyses. SGP analyzed the performance of dual processing model and provided an additional intellectual input. BD wrote the first draft. All authors read and approved the final manuscript.

Acknowledgments

We want to thank to Dr Shira Elqayam of De Montfort University, Leicester, UK for the most helpful comments, in particular to introducing us to a notion of the parallel competitive vs. default-interventionalist dual processing theories and pointing the way how our model can help reconcile these two competing theoretical frameworks.
Presented as a poster at: 14th Biennial European Conference of the Society for Medical Decision Making (SMDM Europe 2012) Oslo, Norway, June 10–12, 2012.
Supported by the US DoA grant #W81 XWH 09-2-0175 (PI Djulbegovic).

Author details

[1]Center for Evidence-based Medicine and Health Outcomes Research, Tampa, FL, USA. [2]Department of Internal Medicine, Division of Evidence-based Medicine and Health Outcomes Research University of South Florida, Tampa, FL, USA. [3]Departments of Hematology and Health Outcomes and Behavior, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA. [4]Department of Mathematics, Indiana University Northwest, Gary, IN, USA. [5]College of Nursing, University of South Florida, Tampa, FL, USA. [6]Division of Clinical Decision Making, Department of Medicine, Tufts Medical Center, Boston, USA. [7]USF Health, 12901 Bruce B. Downs Boulevard, MDC27, Tampa, FL 33612, USA.

Received: 18 June 2012 Accepted: 21 August 2012
Published: 3 September 2012

References

1. Kahneman D: **Maps of bounded rationality: psychology for behavioral economics.** *Am Econ Rev* 2003, **93**:1449–1475.
2. Kahnemen D: *Thinking fast and slow.* New York: Farrar, Straus and Giroux; 2011.
3. Evans JSTBT: *Hypothethical thinking. Dual processes in reasoning and judgement.* New York: Psychology Press: Taylor and Francis Group; 2007.
4. Stanovich KE, West RF: **Individual differences in reasoning:implications for the rationality debate?** *Behavioral and Brain Sciences* 2000, **23**:645–726.
5. Stanovich KE: *Rationality and the Reflective Mind.* Oxford: Oxford University Press; 2011.
6. Croskerry P: **Clinical cognition and diagnostic error: applications of a dual process model of reasoning.** *Adv Health Sci Educ Theory Pract* 2009, **14**(Suppl 1):27–35.
7. Croskerry P: **A universal model of diagnostic reasoning.** *Acad Med* 2009, **84**(8):1022–1028.
8. Croskerry P, Abbass A, Wu AW: **Emotional influences in patient safety.** *J Patient Saf* 2010, **6**(4):199–205.
9. Croskerry P, Nimmo GR: **Better clinical decision making and reducing diagnostic error.** *J R Coll Physicians Edinb* 2011, **41**(2):155–162.
10. Slovic P, Finucane ML, Peters E, MacGregor DG: **Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality.** *Risk Anal* 2004, **24**(2):311–322.
11. Mukherjee K: **A dual system model of preferences under risk.** *Psychol Rev* 2010, **117**(1):243–255.
12. Djulbegovic B, Hozo I, Lyman GH: **Linking evidence-based medicine therapeutic summary measures to clinical decision analysis.** *MedGenMed* 2000, **2**(1):E6.
13. Pauker S, Kassirer J: **Therapeutic decision making: a cost benefit analysis.** *N Engl J Med* 1975, **293**:229–234.
14. Pauker SG, Kassirer J: **The threshold approach to clinical decision making.** *N Engl J Med* 1980, **302**:1109–1117.
15. Djulbegovic B, Desoky AH: **Equation and nomogram for calculation of testing and treatment thresholds.** *Med Decis Making* 1996, **16**(2):198–199.
16. Djulbegovic B, Hozo I, Schwartz A, McMasters K: **Acceptable regret in medical decision making.** *Med Hypotheses* 1999, **53**:253–259.
17. Hozo I, Djulbegovic B: **When is diagnostic testing inappropriate or irrational? Acceptable regret approach.** *Med Decis Making* 2008, **28**(4):540–553.
18. Hozo I, Djulbegovic B: **Will insistence on practicing medicine according to expected utility theory lead to an increase in diagnostic testing?** *Medical Decision Making* 2009, **29**:320–322.
19. Hozo I, Djulbegovic B: **Clarification and corrections of acceptable regret model.** *Medical Decision Making* 2009, **29**:323–324.
20. Kahneman D, Wakker PP, Sarin RK: **Back to Bentham? Explorations of experienced utility.** *Q J Econ* 1997, **112**:375–405.
21. Berwick DM, Hackbarth AD: **Eliminating waste in US health care.** *JAMA: The Journal of the American Medical Association* 2012, **307**(14):1513–1516.
22. Manchikanti L, Falco FJ, Boswell MV, Hirsch JA: **Facts, fallacies, and politics of comparative effectiveness research: part 2 - implications for interventional pain management.** *Pain Physician* 2010, **13**(1):E55–E79.
23. Manchikanti L, Falco FJ, Boswell MV, Hirsch JA: **Facts, fallacies, and politics of comparative effectiveness research: part I. Basic considerations.** *Pain Physician* 2010, **13**(1):E23–E54.
24. Hsee CK, Rottenstreich Y: **Music, pandas and muggers: on the affective psychology of value.** *J Exp Psychol* 2004, **133**:23–30.
25. Rottenstreich Y, Hsee CK: **Money, kisses, and electric shock: On the affective psychology of risk.** *Psychol Sci* 2001, **12**:185–190.
26. Evans JSTBT: *Thinking Twice. Two Minds in One Brain.* Oxford: Oxford University Press; 2010.
27. Evans JSTBT: **Dual-process theories of reasoning: contemporary issues and developmental applications.** *Dev Rev* 2011, **31**:86–102.
28. Edwards W, Miles R Jr, vonWinterfeld D: *Advances in decision analysis. From foundations to applications.* New York: Cambridge University Press; 2007.
29. Simon HA: **Information processsing models of cognition.** *Ann Review Psychol* 1979, **30**:263–296.
30. Reyna VF, Brainerd CJ: **Dual processes in decision making and developmental neuroscience: a fuzzy-trace model.** *Dev Rev* 2011, **31**(2–3):180–206.
31. Reyna VF, Hamilton AJ: **The importance of memory in informed consent for surgical risk.** *Med Decis Making* 2001, **21**(2):152–155.
32. Kahneman D, Tversky A: **The psychology of preferences.** *Sci American* 1982, **246**:160–173.
33. Tversky A, Kahneman D: **Judgements under uncertainty: heuristics and biases.** *Science* 1974, **185**:1124–1131.
34. Djulbegovic B, Hozo II, Fields KK, Sullivan D: **High-dose chemotherapy in the adjuvant treatment of breast cancer: benefit/risk analysis.** *Cancer Control* 1998, **5**(5):394–405.
35. Djulbegovic B, Hozo I: **When should potentially false research findings be considered acceptable?** *PLoS Medicine* 2007, **4**(2):e26.
36. Djulbegovic B, Hozo I: **Linking evidence-based medicine to clinical decision analysis.** *Med Decision Making* 1998, **18**:464. abstract.
37. Zeelenberg M, Pieters R: **A theory of regret regulation 1.0.** *J Consumer Psychol* 2007, **17**:3–18.
38. Zeelenberg M, Pieters R: **A theory of regret regulation 1.1.** *J Consumer Psychol* 2007, **17**:29–35.

39. Tsalatsanis A, Hozo I, Vickers A, Djulbegovic B: A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making. *BMC Medical Informatics and Decision Making* 2010, 10(1):51.
40. Evans JSTBT: On the resolution of conflict in dual process theories of reasoning. *Think Reasoning* 2007, 13(4):321–339.
41. Hammond KR: *Human judgment and social policy. irreducible uncertainty, inevitable error, unavoidable injustice.* Oxford: Oxford University Press; 1996.
42. Duhigg C: *The power of habit: why we do what we do in life and business.* New York: Random House; 2012.
43. Thompson VA, Prowse Turner JA, Pennycook G: Intuition, reason, and metacognition. *Cogn Psychol* 2011, 63:107–140.
44. Brandstatter E, Gigerenzer G: The priority heuristic: making choices without trade-offs. *Psychol Rev* 2006, 113:409–432.
45. Sox HC: Better care for patients with suspected pulmonary embolism. *Ann Intern Med* 2006, 144(3):210–212.
46. Barritt DW, Jordan SC: Anticoagulant drugs in the treatment of pulmonary embolism. A controlled trial. *Lancet* 1960, 1:1309–1312.
47. Segal JB, Eng J, Jenckes MW, Tamariz LJ, Bolger DT, Krishnan JA, Streiff MB, Harris KA, Feuerstein CJ, Bass EB: Diagnosis and treatment of deep venous thrombosis and pulmonary embolism. In *AHRQ Publication No 03-E016.* Washington, DC: Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services; 2003.
48. Linkins L-A, Choi PT, Douketis JD: Clinical impact of bleeding in patients taking oral anticoagulant therapy for venous thromboembolism: a meta-analysis. *Ann Intern Med* 2003, 139(11):893–900.
49. Roy PM, Durieux P, Gillaizeau F, Legall C, Armand-Perroux A, Martino L, Hachelaf M, Dubart AE, Schmidt J, Cristiano M, *et al*: A computerized handheld decision-support system to improve pulmonary embolism diagnosis: a randomized trial. *Ann Intern Med* 2009, 151(10):677–686.
50. Roy P-M, Colombet I, Durieux P, Chatellier G, Sors H, Meyer G: Systematic review and meta-analysis of strategies for the diagnosis of suspected pulmonary embolism. *BMJ* 2005, 331(7511):259.
51. Hull RD: Diagnosing pulmonary embolism with improved certainty and simplicity. *JAMA* 2006, 295(2):213–215.
52. Koreth J, Schlenk R, Kopecky KJ, Honda S, Sierra J, Djulbegovic BJ, Wadleigh M, DeAngelo DJ, Stone RM, Sakamaki H, *et al*: Allogeneic stem cell transplantation for acute myeloid leukemia in first complete remission: systematic review and meta-analysis of prospective clinical trials. *Jama* 2009, 301(22):2349–2361.
53. Cornelissen JJ, Van Putten WL, Verdonck LF, Theobald M, Jacky E, Daenen SM, van Marwijk Kooy M, Wijermans P, Schouten H, Huijgens PC, *et al*: Results of a HOVON/SAKK donor versus no-donor analysis of myeloablative HLA-identical sibling stem cell transplantation in first remission acute myeloid leukemia in young and middle-aged adults: benefits for whom? *Blood* 2007, 109(9):3658–3666.
54. Djulbegovic B: Principles of reasoning and decision-making. In *Decision Making in Oncology Evidence-based management.* Edited by Djulbegovic B, Sullivan DS. New York: Churchill Livingstone, Inc; 1997:1–14.
55. Feinstein AR: The 'chagrin factor' and qualitative decision analysis. *Arch Intern Med* 1985, 145:1257–1259.
56. Detsky AS: Regional variation in medical care. *N Engl J Med* 1995, 333:5890590.
57. Dilts DM: Practice variation: the Achilles' Heel in quality cancer care. *J Clin Oncol* 2005, 23(25):5881–5882.
58. Eddy DM: Variations in physician practice: the role of uncertainty. *Health Aff* 1984, 3(2):74–89.
59. Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder EL: The implications of regional variations in medicare spending. Part 2: health outcomes and satisfaction with care. *Ann Intern Med* 2003, 138(4):288–298.
60. Sirovich BE, Gottlieb DJ, Welch HG, Fisher ES: Regional variations in health care intensity and physician perceptions of quality of care. *Ann Intern Med* 2006, 144(9):641–649.
61. Zhang Y, Baicker K, Newhouse JP: Geographic variation in medicare drug spending. *N Engl J Med* 2010, 363(5):405–409.
62. Hozo I, Djulbegovic B: *Explaining variation in practice: acceptable regret approoach.* Boston: 28th Annual Meeting of the Society for Medical Decision Making; 2006.
63. Shaw NJ, Dear PR: How do parents of babies interpret qualitative expressions of probability? *Arch Dis Child* 1990, 65(5):520–523.
64. Reyna VF: Theories of medical decision making and health: an evidence-based approach. *Med Decis Making* 2008, 28(6):829–833.
65. Djulbegovic B, Paul A: From efficacy to effectiveness in the face of uncertainty: indication creep and prevention creep. *JAMA* 2011, 305 (19):2005–2006.
66. Ofri D: The emotional epidemiology of H1N1 influenza vaccination. *N Engl J Med* 2009, 361(27):2594–2595.
67. Poland GA: The 2009–2010 influenza pandemic: effects on pandemic and seasonal vaccine uptake and lessons learned for seasonal vaccination campaigns. *Vaccine* 2010, 28(Supplement 4(0)):D3–D13.
68. Krantz DH, Kunreuther HC: Goals and plans in decision making. *Judgement and Decision Making* 2007, 2(3):137–168.
69. Zimmerman HJ: *Fuzzy set theory anf its applications.* 3rd edition. Boston: Kluwer; 1996.
70. Society for simulation in healthcare. http://ssih.org/about-simulation (Last accessed: August 27,2012).
71. Kassirer JP, Kopelman RI: *Learning clinical reasoning.* Baltimore: Williams & Wilkins; 1991.

doi:10.1186/1472-6947-12-94
Cite this article as: Djulbegovic *et al.*: Dual processing model of medical decision-making. *BMC Medical Informatics and Decision Making* 2012 12:94.

Tsalatsanis et al. BMC Medical Informatics and Decision Making 2011, 11:77
http://www.biomedcentral.com/1472-6947/11/77

BMC
Medical Informatics & Decision Making

RESEARCH ARTICLE Open Access

Extensions to Regret-based Decision Curve Analysis: An application to hospice referral for terminal patients

Athanasios Tsalatsanis[1*], Laura E Barnes[1], Iztok Hozo[2] and Benjamin Djulbegovic[1,3]

Abstract

Background: Despite the well documented advantages of hospice care, most terminally ill patients do not reap the maximum benefit from hospice services, with the majority of them receiving hospice care either prematurely or delayed. Decision systems to improve the hospice referral process are sorely needed.

Methods: We present a novel theoretical framework that is based on well-established methodologies of prognostication and decision analysis to assist with the hospice referral process for terminally ill patients. We linked the SUPPORT statistical model, widely regarded as one of the most accurate models for prognostication of terminally ill patients, with the recently developed *regret based decision curve analysis* (*regret DCA*). We extend the *regret DCA* methodology to consider harms associated with the prognostication test as well as harms and effects of the management strategies. In order to enable patients and physicians in making these complex decisions in real-time, we developed an easily accessible web-based decision support system available at the point of care.

Results: The web-based decision support system facilitates the hospice referral process in three steps. First, the patient or surrogate is interviewed to elicit his/her personal preferences regarding the continuation of life-sustaining treatment vs. palliative care. Then, *regret DCA* is employed to identify the best strategy for the particular patient in terms of threshold probability at which he/she is indifferent between continuation of treatment and of hospice referral. Finally, if necessary, the probabilities of survival and death for the particular patient are computed based on the SUPPORT prognostication model and contrasted with the patient's threshold probability. The web-based design of the CDSS enables patients, physicians, and family members to participate in the decision process from anywhere internet access is available.

Conclusions: We present a theoretical framework to facilitate the hospice referral process. Further rigorous clinical evaluation including testing in a prospective randomized controlled trial is required and planned.

Background

Introduction

Hospice services have been proven to provide better quality of care to dying patients[1-3] by optimizing pain relief [4,5] and reducing emotional stress [1,6,7]. Furthermore, hospice care is associated with greater patient-family satisfaction[8], is shown to be cost effective[9,10], and most importantly, it has been attributed with increased survival in some patients [11]. Despite these well documented advantages, many terminally ill patients do not reap maximum benefits from hospice care. The fundamental reason for this is related to the **less than optimal and frequently poorly timed referral** of terminally ill patients to hospice [1,12]. As a result, many patients die within a few days of referral, or live many years after the referral was made [13].

According to Medicare regulations, a person should be referred to hospice if his/her "life expectancy (LE) is 6 months or less" [1,14]. Hence, the problem of meaningful referrals relates to the accurate estimation (prognosis) of death within approximately 6 months after evaluation for hospice care. However, statistical models designed to assist physicians in predicting life

* Correspondence: atsalats@health.usf.edu
[1]Center for Evidence-based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL, USA
Full list of author information is available at the end of the article

expectancy (LE), although beneficial [15,16], so far they failed to improve the quality of care at the end of life [17-21].

One such statistical model is SUPPORT (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments), designed to calculate the probability of survival over a period of 180 days [22,23]. Although the SUPPORT model has been well validated [17,22] for prognostication of LE in terminally ill patients, a controlled trial of SUPPORT failed to demonstrate any impact on the overall quality of care for these patients [17,20]. We postulate that this lack of impact may be due to the fact that SUPPORT results, were not linked to any decision methodology that would translate the probability of survival to a hospice referral recommendation. Therefore, the full potential of the model's prognostication power remained unexploited.

In this work, we link the SUPPORT prognostication model with the recently developed decision methodology *regret DCA* [24] to facilitate the hospice referral process. *Regret DCA* relies on regret theory and decision curve analysis [25] to recommend the optimal management strategy for a patient, accounting for the personal attitudes and values of the particular patient or his/her surrogate.

Furthermore, we extend *regret DCA* to incorporate harms and effects of treatment as well as harms associated with the prognostication test to the decision model. The presented methodology is integrated into a comprehensive clinical decision support system developed to facilitate the hospice referral process.

Methods

Dataset

In our analysis, we utilized the entire SUPPORT dataset, both development and validation cohorts. The dataset is presented in detail elsewhere [22]. Medical records of 8,329 seriously ill hospitalized adults are included.

Support model

SUPPORT is a multivariable model designed to estimate probability of survival for seriously ill hospitalized patients over a period of the subsequent 180 days. The model variables include the patient's medical condition compatible with one of eight major diagnostic groupings (Acute Respiratory Failure, Multiple Organ System Failure, Chronic Obstructive Pulmonary Disease, Congestive Heart Failure, Hepatic Cirrhosis, Neurological Coma, Lung or Colon Cancer), the patient's current age, number of days in the hospital before study entry, neurologic status, and 11 physiologic measures recorded on day 3 after study entry [22].

The SUPPORT implementation for the estimation of survival probability is detailed in the appendix. Due to the nature of the hospice referral problem we also express the survival probability in terms of mortality. We can convert the estimated survival probability (SP) (equation A2) to probability of death within 180 days (denoted here as p) using the equation:

$$p = 1 - SP = 1 - P\{T \geq t | \text{disease group} = i\} \quad (1)$$

where SP is the survival probability computed by SUPPORT, $i \in [1,8]$ the patient's disease group, T is the survival time in days, and t is an arbitrary time (typically expressed in days e.g. $t \in [1,180]$).

In terms of accuracy, the SUPPORT model has an area under the receiver-operating characteristics curve (ROC) for prediction of surviving 180 days of 0.79 in the phase I development cohort and 0.78 in the phase II validation cohort [22].

Decision model

Figure 1 depicts the decision tree summarizing the process of hospice referral. The four outcomes and their corresponding utilities (U) shown are:

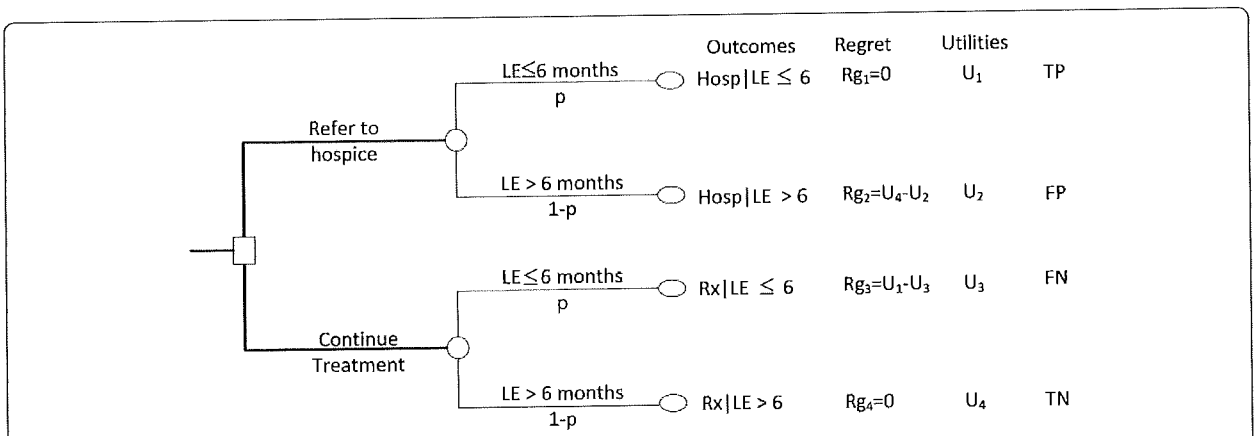


Figure 1 Decision tree for hospice referral. In this figure, p is the probability that a patient's LE is less than or equal to 6 months; 1-p is the probability that a patient's LE is greater than 6 months; U_i: are the utilities associated with each outcome; Rg_i is the regret associated with each outcome.

1. U_1: Refer the patient to hospice and the patient's LE is less than or equal to 6 months ($Hosp|LE \leq 6$).
2. U_2: Refer the patient to hospice and the patient's LE is greater than 6 months ($Hosp|LE > 6$).
3. U_3: Continue treating the patient and the patient's LE is less than or equal to 6 months ($Rx|LE \leq 6$).
4. U_4: Continue treating the patient and the patient's LE is greater than 6 months ($Rx|LE > 6$).

p is the probability associated with the presence of an event (e.g. patient's LE ≤ 6 months) as predicted by the SUPPORT model, 1 - p is the probability associated with the absence of the same event (e.g. patient's LE > 6 months).

As with any decision, one may come to realize that, in retrospect, an alternative decision would have been preferable. This knowledge may bring a sense of loss or regret [26-32]. In this paper, we use this sense of regret to determine the preferences of the decision maker towards alternative management strategies. Specifically, we employ regret theory to estimate the threshold probability, P_t, at which the decision maker (patient, physician, or family member) is indifferent between continuation of treatment vs. hospice referral. Based on the concept of threshold probability, the patient should be referred to hospice if his/her probability of death is greater than or equal to P_t(e.g. $p \geq P_t$), and he/she should continue receiving curative treatment otherwise ($p < P_t$).

The threshold probability is derived as [24]:

$$P_t = \frac{1}{1 + \frac{U_1 - U_3}{U_4 - U_2}} \quad (2)$$

In (2) U_1-U_3 is associated with regret of omission (e.g. the patient was not referred to hospice, instead he/she continued receiving unnecessary treatment) and U_4-U_2 with regret of commission (e.g. the patient was unnecessary referred to hospice instead of continue receiving life-sustaining treatment) [24].

To elicit the decision maker's regret, and therefore threshold probability, we utilize the DVAS (Dual Visual Analogue Scale) method [24]. One visual analogue scale is used to capture the regret associated with failing to refer the patient to hospice (e.g. continue unnecessary treatment) and the second scale to measure the regret associated with unnecessary hospice referral (e.g. failing to provide life-sustaining treatment) (Figure 2).

Elicitation of threshold probability can be achieved through a set of questions such as:

1. *On the scale 0 to 100, where 0 indicates no regret and 100 indicates the maximum regret you could feel, how would you weigh the level of your regret if you were not referred to hospice but instead you continued receiving unnecessary treatment? That is, how much would you regret if you did not reap the benefits of hospice care?* *Note that this value corresponds to U_1 - U_3.
2. *On the scale 0 to 100, where 0 indicates no regret and 100 indicates the maximum regret you could feel, how would you weigh the level of your regret, if you were referred to hospice instead of continue receiving necessary life-sustaining treatment? That is, how much would you regret if you sustained harms from hospice care?* *Note that this value corresponds to U_4 - U_2.

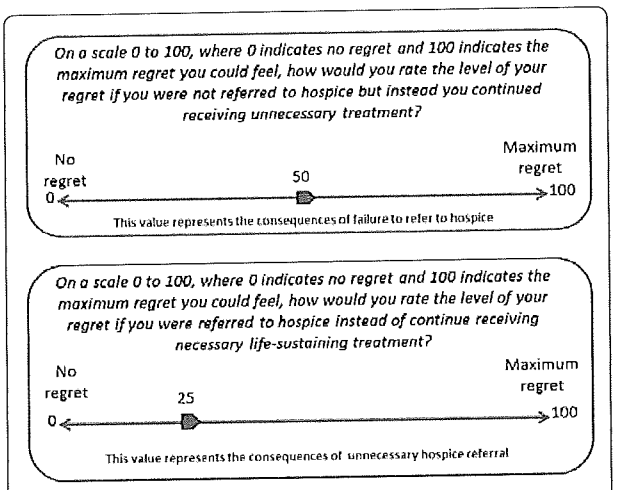


Figure 2 DVAS (Dual Visual Analogue Scales). The DVAS are used for the elicitation of the decision maker's threshold probability.

For example, suppose that the patient - who is aware of his/her terminal condition- answers 50 and 25 to the questions 1 and 2 respectively. This means that the patient considers 50/25 = 2 times worse not to be referred to hospice when necessary than receiving an unnecessary hospice referral. The threshold probability for this patient is (equation 2)

$$p_t = \frac{1}{1 + \frac{U_1 - U_3}{U_4 - U_2}} = \frac{1}{1 + \frac{50}{25}} = 0.33 \text{ or } 33\% .$$

Regret DCA and extensions

The clinical problem we face in the situation of hospice referral is how to use reasonably accurate predictions of death, p, coupled with the patient's preferences (as expressed in terms of threshold probability, P_t) to arrive at the optimal decision for a specific individual. The problem is decomposed into three strategies: (1) act based on the prediction model (SUPPORT) (e.g. refer to hospice if $p \geq P_t$ and continue treating otherwise), (2)

refer all patients to hospice, and (3) continue current treatment for all patients (i.e. refer no patients to hospice).

Each of these strategies may inflict physiological and/or psychological damages to the patient. Specifically, a patient may suffer harms due to a treatment strategy (e.g. adverse effects) or harms due to the prognostication test (e.g. a test requiring invasive procedure). We express these harms as loss in utility associated with actions we may undertake. To that end, we define H_{Rx}, H_{Hosp} and H_{te} as the utility losses due to harms of the treatment, hospice, and prognostic test, respectively.

Figure 3 presents the decision tree describing the overall hospice referral problem. $p = P(D+)$ is the probability that the patient's LE is less than or equal to 6 months as estimated by the prediction model (SUPPORT);$1 - p = P(D-)$ is the probability that the patient's LE is greater than 6 months, and U_i, $i \in [1,4]$, are the utilities corresponding to each of the decision model outcomes (detailed in the previous section). The variables *Hosp* and *Rx* correspond to referring a patient to hospice and continuing current curative treatment, respectively. *Rg*, is the regret associated with an action, e.g. *Rg(Hosp, D-)* is the regret one may feel if the patient was referred to hospice when his/her LE was greater than 6 months. Finally, *te* designates that the patient received a prognostication test.

Considering the decision tree in Figure 3 we can compute the expected regret associated with each decision in terms of the utilities of each possible outcome as follows (detailed derivation is presented in the Appendix):

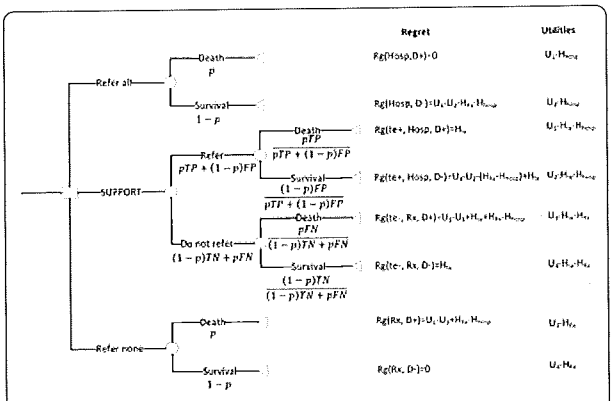


Figure 3 Decision tree describing the overall hospice referral process. In this figure p = P (D +): probability the patient's LE is less than or equal to 6 months; 1 - p = P (D -): probability the patient's LE is greater than 6 months; U_i, i ∈ [1,4]: the utilities corresponding to each of the decision model outcomes; Hosp: hospice referral; Rx treatment continuation; Rg: regret associated with an action; H_{Rx} : utility losses due to harms of treatment; H_{Hosp}: utility losses due to harms of hospice; H_{te}: utility losses due to harms of the prognostic test (SUPPORT).

$$ERg[Hosp] = (1-p) * (1 - RRR_{Hosp}) * \frac{P_t}{1-P_t} \quad (3)$$

$$ERg[Rx] = p * (1 - RRR_{Rx}) \quad (4)$$

$$\begin{aligned} &ERg[SUPPORT] = \\ &\left(1 - RRR_{Hosp} * (\#TP/n + \#FP/n)\right. \\ &\left. -RRR_{Rx} * (\#FN/n + \#TN/n)\right) \\ &*\frac{H_{te}}{U_1 - U_3 + H_{Rx} - H_{Hosp}} \\ &+(1 - RRR_{Hosp}) * \frac{\#FP}{n} * \frac{P_t}{1 - P_t} \\ &+(1 - RRR_{Rx}) * \frac{\#FN}{n} \end{aligned} \quad (5)$$

In addition to harms, equations 3, 4 and 5 incorporate the effects of treatment and hospice care using measures of *Relative Risk Reduction:* and RRR_{Rx} RRR_{Hosp} respectively. The values for these measures are treatment specific and can be acquired from the literature. We have incorporated hospice effects because a recent study [11] has shown that early palliative care for patients with metastatic non-small cell lung cancer could increase survival. The variables *TP, FP, FN, TN* are related to the prognostic capability of the SUPPORT model (see appendix for detailed derivation) [24].

Since the regret of omission and regret of commission have been generalized to include effects and harms related to management strategies and testing, the function of threshold probability (equation 2) becomes:

$$P_t = \frac{1}{1 + \frac{U_1 - U_3 + H_{Rx} - H_{Hosp}}{U_4 - U_2 - H_{Rx} - H_{Hosp}}} \quad (6)$$

Where U_1 - U_3 + H_{Rx} -H_{Hosp} corresponds to the regret associated with not referring the patient to hospice when necessary, and U_4 - U_2 - H_{Rx} - H_{Hosp} corresponds to the regret associated with unnecessary hospice referral.

Choosing the optimal strategy

The optimal strategy is selected as the one which will bring the least amount of regret. The *regret DCA* algorithm expresses the regret associated with each strategy in terms of threshold probability and is implemented as follows [24]:

1. Select a value for threshold probability.
2. Assuming that patients should be referred to hospice if $p \geq P_t$ and should continue current treatment otherwise, compute #TP and #FP for the prediction model.
3. Calculate the *ERg(SUPPORT)*using equation 5.

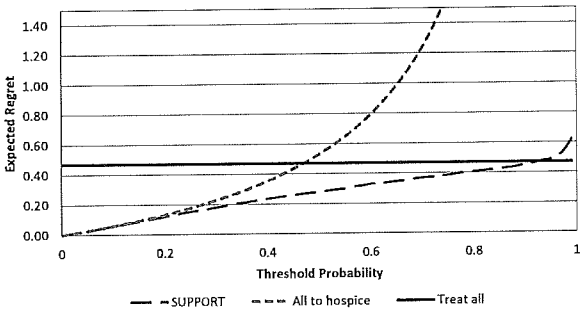


Figure 4 Decision curves for hospice referral. In this figure, RRR_{Hosp} = 0, RRR_{Rx} = 0, H_{Rx} = H_{Hosp} = H_{te} = 0. At threshold probability equal to 10%, the optimal decision is refer the patient to hospice; at 40% the optimal decision is to use the SUPPORT model.

4. Calculate *ERg(Rx)* using equation 4.
5. Compute the *ERg(Hosp)*using equation 3.
6. Repeat steps 1 - 6 for a range of threshold probabilities.
7. Graph each expected regret function calculated in steps 3-5 against each threshold probability.

At each threshold probability, **the action with the lowest value of expected regret corresponds to the most desired action**. For example, in Figure 4, at a threshold probability equal to 10% (e.g. the patient considers 9 times worse not to be referred to hospice when necessary than to receive an unnecessary hospice referral), the optimal strategy is to refer the patient to hospice.

Figures 5, 6 and 7 depict the regret associated with alternative decision strategies as they relate to different

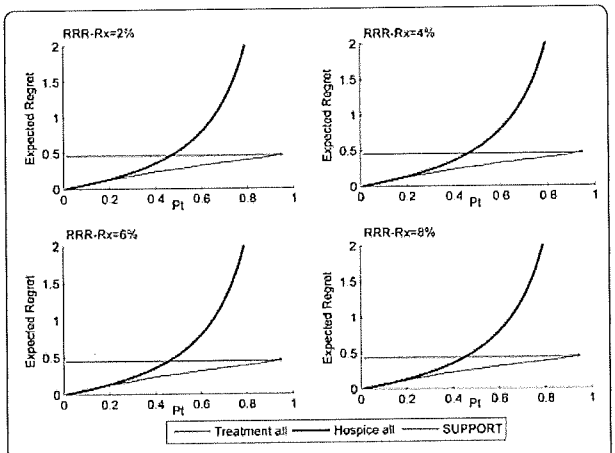


Figure 5 Decision curves as a function of RRR_{RX}. In this figure, RRR_{Hosp} = 0, RRR_{Rx} = 2 to 8%, H_{Rx} = H_{Hosp} = H_{te} = 0. As the effect of treatment increases, the regret associated with treating all patients and with the SUPPORT model slightly decreases. The strategy of using the SUPPORT model to refer a patient to hospice is the action with the least amount of regret for the wider range of threshold probabilities.

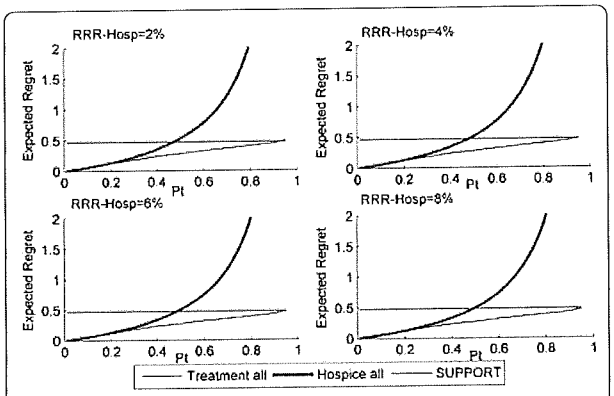


Figure 6 Decision curves as a function of RRR_{Hosp}. In this figure, RRR_{Hosp} = 2% to 8%, RRR_{Rx} = 0, H_{Rx} = H_{Hosp} = H_{te} = 0. As the effect of hospice care increases, the regret associated with hospice and with the SUPPORT model slightly decreases. As previously, the strategy of using the SUPPORT model to refer a patient to hospice is the action with the least amount of regret for a wide range of threshold probabilities.

values of hospice effectiveness (Figure 5), treatment effectiveness (Figure 6), and harms due to the prognostication test (Figure 7). As expected, when the harms due to the prognostication test are increased, then the area of threshold probability at which the prognostication model is the optimal decision is reduced (Figure 5). Even though, it is not expected that the SUPPORT model will actually create harms, at least physiological, to the patient, this is not always the case for other diagnostic tests that may be more invasive (e.g. screening for prostate cancer).

As can be seen from Figures 5, 6 and 7 the optimal decision is derived by the SUPPORT model for a rather wide range of threshold probabilities. Therefore, it appears that the SUPPORT model is the superior

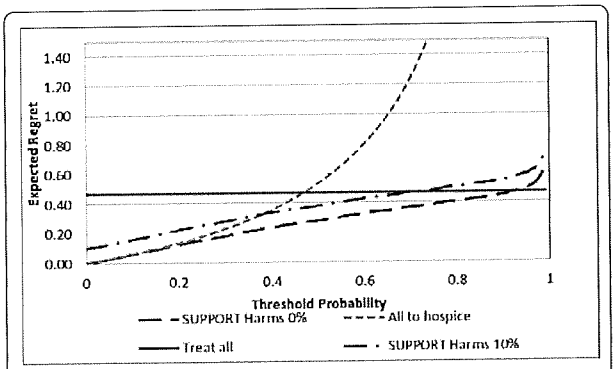


Figure 7 Decision curves as a function of H_{te}. In this figure, RRR_{Hosp} = 0, RRR_{Rx} = 0, H_{Rx} = H_{Hosp} = 0 and H_{te} = 0% and 10%. As the harms associated to the prediction test increase, so does the expected regret of utilizing the SUPPORT model for hospice referral. Increasing the harms due to treatment or due to hospice care does not have an effect on the decision curves.

strategy for the vast majority of decision makers, regardless the effects of the alternative management strategies. However, since the threshold probability expresses the personal preferences of a particular decision maker, it is not unusual for specific patients to have smaller or greater threshold probability values than the majority of decision makers. This is the power of the proposed methodology, which allows for decision making at the individual level. For example, if the decision maker presents a threshold probability greater than ≈92%, the optimal decision would be to continue life-sustaining treatment even if it is deemed not to be effective(Figure 4). Similarly, for small values of threshold probability, the desired action would be to refer the patient to hospice.

Decision Support System

As our theoretical discussion highlighted, decisions about life and death are complex and difficult at both the emotional and cognitive level. Therefore, it is not surprising that the SUPPORT model originally failed to improve the quality of care for terminally ill patients despite its reasonable accuracy in prediction of probability of survival [17,20]. Any attempt to focus on a single dimension of the complex hospice referral process is not likely to succeed. An accurate prognostic model is only the first step. Having the apparatus to take into account trade-offs associated with the hospice referral decision while taking into consideration the patients' preferences represent further necessary steps to improve the care of terminally ill patients. In addition, we hypothesize that the SUPPORT intervention failed because it was not available at the point of care in real time. This is because the most desired outcomes are best achieved when decision-making occurs in real-time, at the point of care [33,34].

To facilitate the decision making process for the hospice referral at bedside, we propose a web-based clinical decision support system (CDSS) that computes the probabilities of survival and death for individual patients using the SUPPORT model, elicits personal preferences from patients and/or physicians, and utilizes *regret DCA* to suggest the optimal decision for a particular patient.

Features

Access

Our goal is to develop a CDSS that can be accessed by everyone and from anywhere regardless the operating system one uses. At the same time, it is desirable to develop a system that can eventually be integrated with various healthcare providers' electronic medical records (EMR). We concluded that a web-based implementation would fulfil such requirements.

Data storage

The CDSS performs the required computations without retaining or transmitting sensitive and identifiable information.

Results

In this section we present a prototype of the CDSS, developed to demonstrate the applicability of our theoretical framework for hospice referral. Each subsection describes the results of the methods shown in the previous section in conjunction with the description of the corresponding module. Figure 8 depicts the logical diagram that outlines the operation of the CDSS. Briefly, the operation begins by interviewing the patient or surrogate to elicit his/her threshold probability. Based on the value of threshold probability equation, the optimal strategy for the particular patient is derived (e.g. refer to hospice, continue treatment, or use of prediction model). If the optimal strategy is to follow the

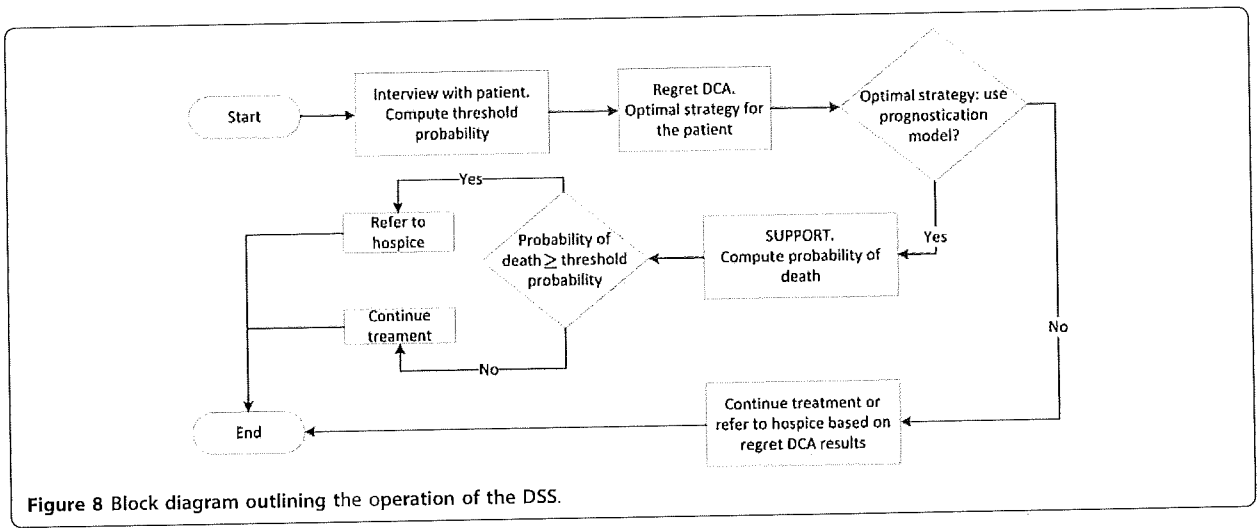


Figure 8 Block diagram outlining the operation of the DSS.

prediction model (SUPPORT), then using the equations A1, A2 and 1, the probabilities of survival and death are computed for the particular patient. The probability of death is then contrasted with the patient's threshold probability and the optimal decision is derived (refer to hospice, or continue treatment). At each step described, the patient selects the level of information he/she wishes to be exposed to. For example, the patient may not wish to know his/her threshold probability or probability of death. Instead he/she wishes to know only the optimal decision regarding his/her condition.

General implementation details

The proposed CDSS is a web-based application residing on the USF Health servers. The web address is http://health.usf.edu/research/ebm/decisionaids.htm. It has been developed based on the Adobe® ColdFusion® application technology and the interface has been designed using html and JavaScript programming languages. The hardware and software requirements from the user's point of view are modest. The system runs on any contemporary computer with net browsing capabilities. However, at this stage the CDSS is not optimized for use with handheld devices. The CDSS consists of 3 different modules as described below.

Elicitation of threshold probability module

The threshold elicitation module consists of the dual visual analogue scales, used to weigh the patient's regret in the case of wrong decisions. Each scale has 100 points where 0 corresponds to no regret and 100 to maximum regret. Depending on the role of the decision maker (e.g. patient/surrogate or physician) two different sets of questions are displayed. These questions are designed to capture the regret of omission and the regret of commission. For the remainder of this paper, we assume that the decision maker is the patient. As in pain scales [35], each visual analogue scale uses facial expressions to graphically represent variations in regret (Figure 9). A summary of the decision maker's preferences is presented for final verification. The threshold probability for the particular patient is derived using equation 2, however is not displayed until the decision maker requests it.

Decision module

The decision module utilizes the decision maker's threshold probability and the regret DCA methodology to derive the optimal decision. For example, the preferences of the patient depicted in Figure 9, correspond to a threshold probability equal to 29%. From Figure 4 the

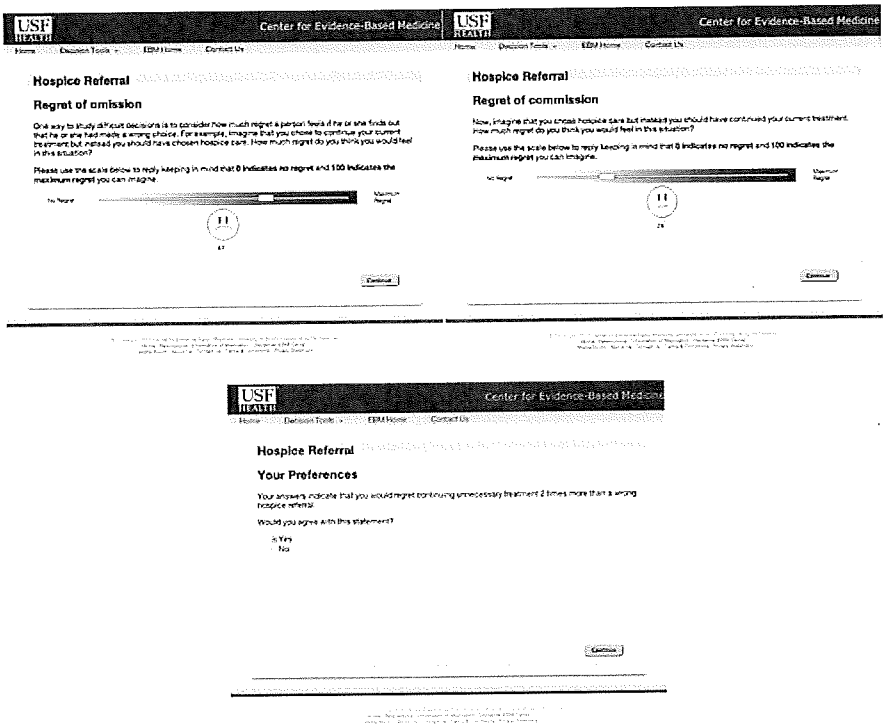


Figure 9 Elicitation of threshold probability. The user (patient/surrogate/physician) weighs the two alternative management strategies in terms of regret.

strategy that will bring the least amount of regret is to use the prognostication model (SUPPORT) for the hospice referral recommendation. In this case, the decision module initiates the SUPPORT module.

SUPPORT module

If the optimal strategy derived by the decision module is to utilize the prognostication model, the SUPPORT module is enabled. This module (Figure 10) is used to compute the probability of death for the particular patient based on the SUPPORT prognostication model. Currently, the user inserts all required information to the CDSS. In the future, this information will be captured automatically from the health care provider's electronic medical records system. Data validation restrictions have been imposed to protect the integrity of the collected data.

Once the values of all available variables have been inserted in the corresponding cells, the patient's life expectancy and probabilities of survival and death are computed. The decision module is employed again to display the optimal recommendation.

Decision justification module

The decision justification module explains in detail and at the user's request the reasons that led to a particular recommendation (Figure 11). It contains information regarding the decision maker's threshold probability, the optimal strategy associated with the threshold probability and the patient's probability of death (if applicable). Since people often misinterpret probabilities [36], we complement the results presented in terms of probabilities using frequency format (Figure 11). The latter format is currently considered the best way to represent favorable and unfavorable facts regarding medical interventions [37]. The justification module is highly technical and should only be reviewed by decision makers who wish to know more about their or their patient's condition.

Case Study

Figure 12 summarizes the decision process for a patient whose information is simulated in Figures 10, 11 and 12. The probability of death and the threshold probability of this patient have been computed as 85% and 29% respectively. At a 29% threshold probability, the optimal strategy is to use the prediction model for hospice referral (Figure 4). Therefore, since $p > P_t$ the patient should be referred to hospice. For completeness, all possible decision routes are depicted in Figure 12. The route corresponding to the specific simulated patient is shown using bold arrows.

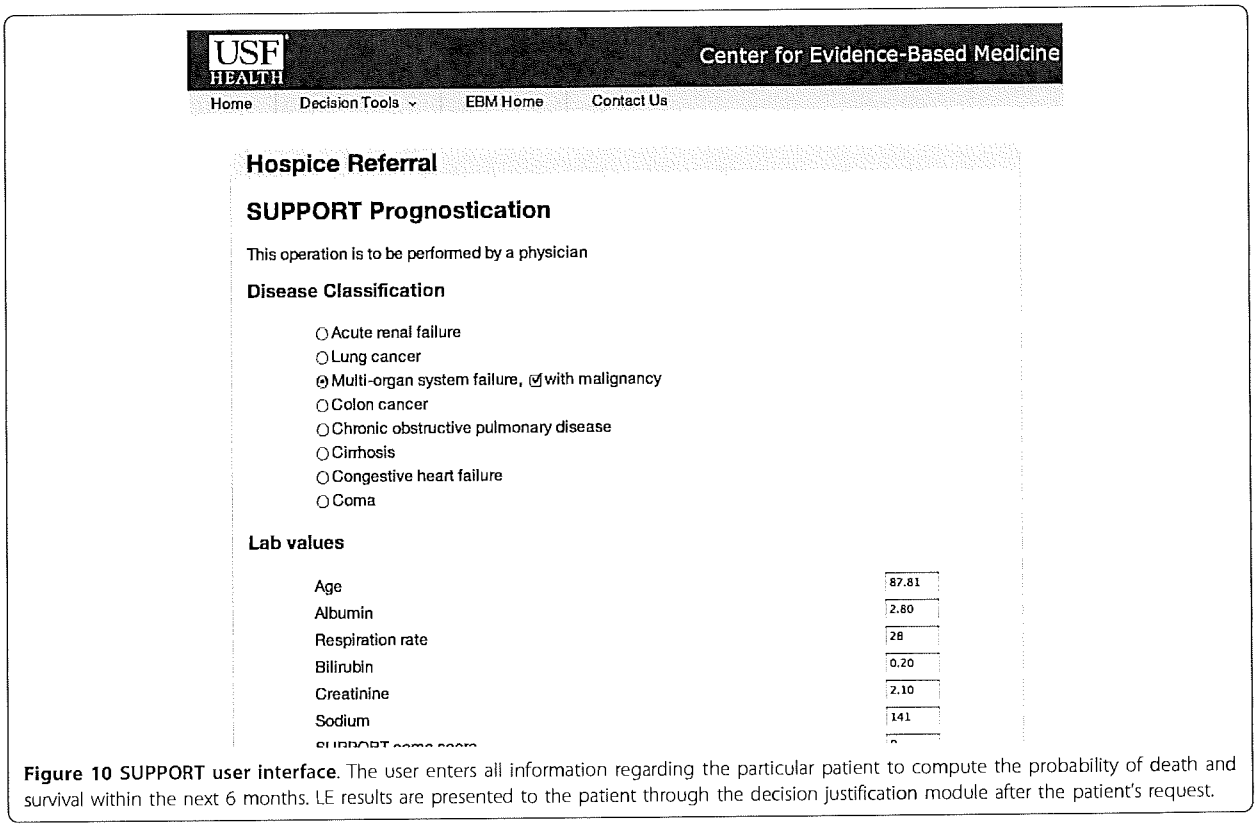


Figure 10 SUPPORT user interface. The user enters all information regarding the particular patient to compute the probability of death and survival within the next 6 months. LE results are presented to the patient through the decision justification module after the patient's request.

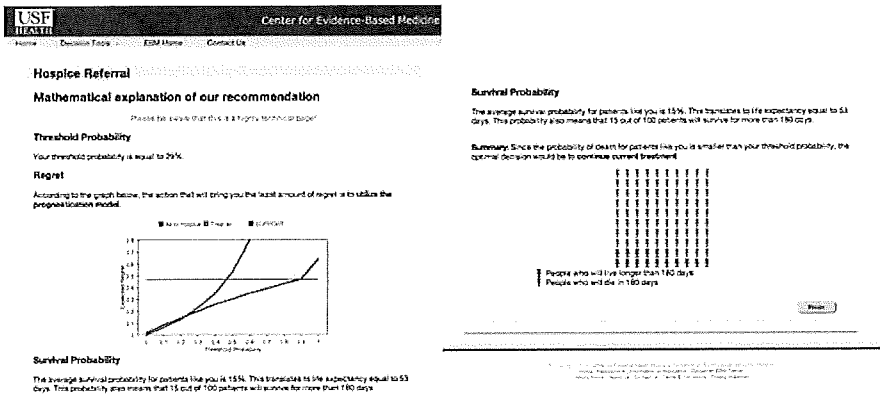


Figure 11 Justification of the hospice referral recommendation. The particular patient depicted has 29% threshold probability at which the optimal strategy is derived by the SUPPORT model. The patient has 85% probability of death in the next 180 days. Therefore, the optimal decision is to be referred to hopsice.

Discussion

In this article we describe both the theory and application behind a hospice referral clinical decision support system. To the best of our knowledge, this is the first CDSS that integrates two well established methodologies, one for prognostication (SUPPORT) and the other for decision making (*regret DCA*), to assist with the hospice referral decision-making process.

The recently developed *regret DCA* incorporates the decision maker's preferences towards alternative management strategies from the perspective of regret theory in terms of threshold probability. Such an approach promotes personalized patient care. We anticipate that the regret-based approach is more appropriate for the hospice referral process than other preference elicitation techniques, due to the nature of the problem where there are really no optimal options available- the optimal decision can be only considered as the one with the least regret.

Modern cognitive theories increasingly focus on the so called dual-processing theory in which both intuition (system 1) and analytical, deliberative process (system 2) are important for balancing risks and benefits in the decision-making process [38]. We believe that rational decision-making should take into account both formal principles of rationality and human intuition about good decisions[24,39,40]. One way to accomplish this is to use regret, a cognitive emotion, to serve as the link between systems 1 and 2 [24]. By taking into account the consequences of our actions as well as the circumstances under which we can live with our mistakes we anticipate that the goal of reconciling the formal

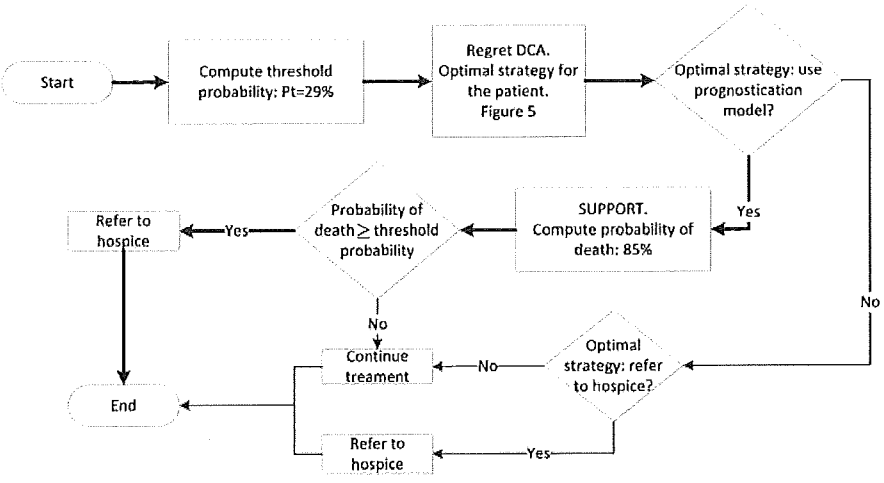


Figure 12 Block diagram summarizing the decision process. The bold route corresponds to the patient simulated in figures 10, 11 and 12.

Table 1 Values of Survival (S) as described in equation A2 for different disease types and varying survival times

t	$S_{ARF/MOSF}$	$S_{COPD/CHP/Cirrhosis}$	S_{Coma}	S_{Cancer}
0	0.994	0.998	0.993	0.993
30	0.691	0.889	0.630	0.578
60	0.601	0.837	0.609	0.407
90	0.562	0.800	0.581	0.264
120	0.532	0.772	0.569	0.190
150	0.508	0.751	0.551	0.135
177	0.493	0.733	0.545	0.108

principles of rationality and human intuitions about good decisions can be met [24,29,39,40]. This is particularly true in the situation of terminally ill patients.

Our web-based CDSS reflects modern cognitive theories to facilitate integration of the decision-making ingredients necessary for hospice referral decisions. The CDSS encapsulates all required information for the hospice referral process into a flexible software that can be used at bedside. Obviously, hospice referral decisions are complex and must be exercised with full compassion and deliberation. We advise against the use of our system as an automatic decision making tool that by-passes important personal interactions between the patient and his/her physician. It is important to stress that the elicitation of the threshold probability as described herein reflects the belief (also captured in recent legislation [41]) that patients and their families want to be told the "truth" about the patients' terminal sickness [22,41,42] and that physicians have ethical obligations to share this information with patients and their families [41,42]. Our system should be understood as an aid to facilitate decisions in terminal phases of patient lives.

Our approach has limitations as well. The main limitation of the proposed system remains the complexity of the SUPPORT model. Currently, the system still requires manual entry of data. In addition, failure to enter all data can jeopardize the accuracy of prediction and therefore, the decision process. To cope with this limitation, we plan to integrate our system into various health providers' EMRs. Based on each EMR, specifically designed queries will be used to retrieve lab values and patient demographics to be fed automatically into our system; a process that will reduce the amount of missing values and input errors.

The second limitation of the proposed system is that empirical data are not available to assess how the system actually works in practice. While we plan to undertake empirical testing of the system described here, we believe that a strong theoretical underpinning will enable better hospice referral decisions even in the current form. This is because our system will essentially operationalize the decision-making process, which is supposed to occur in every day practice. Nevertheless, we need to firstly, identify the system's feasibility in real life settings and ultimately, if it appears to be usable and assessed favourably by all those involved in the hospice-referral decision-making process, to test it in randomized controlled trials against traditional care.

Our future plans include both empirical testing and implementation of multiple additional prognostication models which will be used in parallel to assess optimal decisions regarding hospice referral and take advantage of the regret DCA methodology. We anticipate that for a different range of threshold probabilities these models may perform better than the SUPPORT model. Furthermore, our intent is to develop a separate version of our CDSS optimized for mobile devices.

Conclusions

In this work we have presented the theoretical framework, accompanied by the associated CDSS, to facilitate end of life care decisions. Our work combines the prognostication power of the SUPPORT model, the simplicity of the DVAS methodology in eliciting people's preferences and the effectiveness of regret DCA at evaluating alternative management strategies to resolve the dilemma of choosing traditional vs. palliative care for patients at terminal stages. A clinical evaluation of the CDSS is planned.

Appendix

Support implementation

SUPPORT is implemented in two steps. First, the SUPPORT physiology score is computed based on equation A1 [22].

$$
\begin{aligned}
SPS = {} & 259.9\{ARF/MOSF\} + 263.4\{COPD/CHF\} \\
& + 241.4\{Cirrhosis/Coma\} \\
& + 281.5\{Lung/ColonCancer\} \\
& - 0.06174\min(PaO_2/FiO_2, 225) \\
& - 0.6316\min(MeanBP, 60) \\
& + 1.0205WBC - 0.3676(WBC - 8)_+ \\
& - 0.5631(WBC - 11)_+ + 0.2691\min(Alb, 4.6) \\
& + 0.2312Aresp - 2.362Temp + 1.326(Temp - 36.6)_+ \\
& + 2.473(Temp - 38.3)_+ - 1.579 \times 10^{-}{}^{1}HR \\
& + 9.770 \times 10^{-5}(HR - 55)_+^3 - 2.189 \times 10^{-4}(HR - 80)_+^3 \\
& + 1.518 \times 10^{-4}(HR - 110)_+^3 \\
& - 3.062 \times 10^{-5}(HR - 149)_+^3 + 0.9763Bil \\
& - 0.7481(Bil - 7)_+ - 6.8761Cr \\
& + 11.6058(Cr - 0.600)_+^3 - 21.8413(Cr - 1.000)_+^3 \\
& + 10.3574(Cr - 1.500)_+^3 - 0.1219(Cr - 5.399)_+^3 \\
& - 0.6167096Na + 0.0021118(Na - 128)_+^3 \\
& - 0.0036730(Na - 135)_+^3 + 0.0006126(Na - 139)_+^3 \\
& + 0.0009486(Na - 148)_+^3 \\
& - 6.278\{COPD/CHF\} \times \min(Alb, 4.6) \\
& - 11.45\{Lung/ColonCancer\} \times \min(Alb, 4.6) \\
& + \{ARF/MOSF\}[-2.3549WBC \\
& + 2.7494(WBC - 8)_+ - 0.4638(WBC - 11)_+]
\end{aligned}
\tag{A1}
$$

where:*Alb*: albumin; *Aresp*: APACHE III respiration score; *Bil*: bilirubin; *Cr*: Creatinine; *Na*: sodium; PaO_2: partial pressure oxygen in arterial blood; *MeanBP*: mean arterial blood pressure; *WBC*: white blood cell count in thousands; *Temp*: temperature in Celsius; *HR*: heart rate per minute; *ARF*: Acute respiratory failure; *MOSF*: Multiple organ failure; *Cirrhosis*: Cirrhosis; *Coma*: Coma; *Lung*: Lung cancer; *ColonCancer*: Colon cancer; *COPD*: Chronic obstructive pulmonary disease; *CHF*: Congestive heart failure. Also:

$$\{disease\ group\} = \begin{cases} 1, & \text{if patient in the disease group} \\ 0, & \text{otherwise} \end{cases}$$

$$(x)_+ = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

$WBC = 9$, if $WBC < 9$ and $\{disease\ group\} \neq ARF/MOSF$

$WBC = 40$, if $WBC > 40$

$Cr = 15$, if Cr >15

The second step in implementing the SUPPORT model is to calculate the probability of survival for the individual patient based on equation A2 [22].

$$P\{T \geq t | \text{disease group} = i\} = S_i(t)^{e^{X_b}} \quad \text{(A2)}$$

where T: survival time in days; t: arbitrary time; S described in Table 1[22] and

$$\begin{aligned} X_{\hat{b}} &= -3.652 + 0.8356\{CHF\} + 0.9257\{Cirrhosis\} \\ &+ 0.6287\{LungCancer\} \pm 1.1803\{MOSFw/Malig\} \\ &+ 0.01434 Scoma \pm 0.01935 Age + 0.2413 Cancer \\ &- 1.863\left[Hday + 3.4\right]^{-1} + 0.08121 SPS \\ &+ Age[0.015261\{COPD/CHF/Cirrhosis\} \\ &+ 0.009047\{Coma\} - 0.008294\{Cancer\}] \\ &+ Age[-0.012498\{CHF\} - 0.004578\{Cirrhosis\} \\ &- 0.001435\{LungCancer\} \\ &- 0.013891\{MOSFw/Malig\} \end{aligned}$$

where *Scoma*: SUPPORT coma score (0-100); *MOSFw/Malig*: Multiple organ failure with malignancies; *Hday*: day in hospital when qualified for study; *Cancer*: Cancer by comorbidity or primary disease category (0 = no; 1 = present; 2 = metastatic) [22].

Derivation of the Expected Regret functions

As outlined in the Introduction, seriously and terminally ill patients may reap a number of benefits by the hospice program. Nevertheless, after enrollment into hospice, the patient (or the family, or the physician) may feel that this was a wrong decision, and subsequently may regret it. Similarly, the patient may feel regret for the treatment that he/she continues to receive because it is unnecessary, inappropriate, and/or harmful. Figure 3 represents our hospice decision tree in terms of regret from which we can compute the expected values of regret associated with each strategy as follows:

$$\begin{aligned} &ERg[Hosp] = \\ &(1-p) * (U_4 - U_2 - H_{Rx} - H_{Hosp}) \end{aligned} \quad \text{(A3)}$$

$$ERg[Rx] = p * (U_1 - U_3 + H_{Rx} - H_{Hosp}) \quad \text{(A4)}$$

$$\begin{aligned} &ERg[SUPPORT] = p * TP * H_{te} \\ &+ (1-p) * FP \\ &*(U_4 - U_2 - (H_{Rx} - H_{Hosp}) + H_{te}) \\ &+p * FN * (U_1 - U_3 + (H_{Rx} - H_{Hosp}) + H_{te}) \\ &+(1-p) * TN * H_{te} \end{aligned} \quad \text{(A5)}$$

The variables *TP*, *FP*, *TN*, *FN* are related to the probabilities $P\ (p \geq P_t \cap D+)$, $P\ (p \geq P_t \cap D-)$, $P\ (p < P_t \cap D-)$ and $P\ (p < P_t \cap D+)$ respectively, and are estimated as follows:

- $P\ (p \geq P_t \cap D+) \approx$ the number of patients who will die within 6 months and for whom the prognostic probability is greater than or equal to P_t (with #TP = number of patients with true positive results, $P(p \geq P_t \cap D+) \approx \frac{\#TP}{n}$, where n is the total number of patients in the study).
- $P\ (p \geq P_t \cap D-) \approx$the number of patients who will survive for longer than 6 months and for whom the prognostic probability is greater than or equal to P_t (with #FP = number of patients with false positive results, $P(p \geq P_t \cap D-) \approx \frac{\#FP}{n}$).
- $P\ (p < P_t \cap D+) \approx$the number of patients who will die within 6 months and for whom the prognostic probability is less than P_t (with #FN = number of patients with false negative results, $P(p < P_t \cap D+) \approx \frac{\#FN}{n}$).
- $P\ (p < P_t \cap D-) \approx$the number of patients who will survive for longer than 6 months and for whom the prognostic probability is less than P_t (with #TN = number of patients with true negative results, $P(p < P_t \cap D-) \approx \frac{\#TN}{n}$).

To incorporate the effects of alternative treatments (e.g. treatment and hospice care) in equations A3-A5 we use the *Relative Risk Reduction* reported in literature for each strategy as follows:

$$ERg[Hosp] = (1-p) * (1 - RRR_{Hosp}) * (U_4 - U_2 - H_{Rx} - H_{Hosp}) \quad (A6)$$

$$ERg[Rx] = p * (1 - RRR_{Rx}) * (U_1 - U_3 + H_{Rx} - H_{Hosp}) \quad (A7)$$

$$\begin{aligned} ERg[SUPPORT] = {} & p * (1 - RRR_{Hosp}) * TP * H_{te} \\ & + (1-p) * (1 - RRR_{Hosp}) * FP \\ & * (U_4 - U_2 - (H_{Rx} - H_{Hosp}) + H_{te}) \\ & + p * (1 - RRR_{Rx}) * FN \\ & * (U_1 - U_3 + (H_{Rx} - H_{Hospe}) + H_{te}) \\ & + (1-p) * (1 - RRR_{Rx}) * TN * H_{te} \end{aligned} \quad (A8)$$

Since $TP + FN = 1$ and $FP + TN = 1$, we have:

$$p * TP + (1-p) * FP + p * FN + (1-p) * TN = p + (1-p) = 1$$

Therefore, equation A8 becomes:

$$\begin{aligned} ERg[SUPPORT] = {} & (1 - p * RRR_{Hosp} * TP - (1-p) * RRR_{Hosp} * FP \\ & - p * RRR_{Rx} * FN - (1-p) * RRR_{Rx} * TN) * H_{te} \\ & + (1-p) * (1 - RRR_{Hosp}) * FP \\ & * (U_4 - U_2 - (H_{Rx} - H_{Hosp})) \\ & + p * (1 - RRR_{Rx}) * FN \\ & * (U_1 - U_3 + (H_{Rx} - H_{Hosp})) \end{aligned} \quad (A9)$$

Scaling the equations A3, A4 and A9 with the quantity $(U_1 - U_3 + H_{Rx} - H_{Hosp})$ and replacing the expression $\frac{U_4 - U_2 - (H_{Rx} - H_{Hosp})}{U_1 - U_3 + H_{Rx} - H_{Hosp}}$ with $\frac{P_t}{1-P_t}$, we derive the final equations for the expected regret (equations 3, 4, and 5).

Acknowledgements
This work is supported by the Department of Army grant #W81 XWH 09-2-0175 (PI Djulbegovic).

Author details
[1]Center for Evidence-based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL, USA. [2]Department of Mathematics, Indiana University Northwest, Gary, IN, USA. [3]H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA.

Authors' contributions
All authors contributed equally to this work. All authors have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 26 August 2011 Accepted: 23 December 2011
Published: 23 December 2011

References
1. Christakis NA: **Death foretold. Prophecy and prognosis in medical care.** The University of Chicago Press; 2001.
2. Bulkin W, Lukashok H: **Rx for dying: the case for hospice.** *N Engl J Med* 1988, **318(6)**:376-378.
3. Christakis NA, Iwashyna TJ: **Spousal illness burden is associated with delayed use of hospice care in terminally ill patients.** *J Palliat Med* 1998, **1(1)**:3-10.
4. Goldberg RJ, Mor V, Wiemann M, Greer DS, Hiris J: **Analgesic use in terminal cancer patients: report from the National Hospice Study.** *J Chronic Dis* 1986, **39(1)**:37-45.
5. Greer DS, Mor V: **An overview of National Hospice Study findings.** *J Chronic Dis* 1986, **39(1)**:5-7.
6. Moinpour CM, Polissar L: **Factors affecting place of death of hospice and non-hospice cancer patients.** *Am J Public Health* 1989, **79(11)**:1549-1551.
7. Kane RL, Klein SJ, Bernstein L, Rothernberg R, Wales J: **Hospice role in alleviating the emotional stress of terminal patients and their families.** *Med Care* 1985, **23(3)**:189-197.
8. Dawson NJ: **Need satisfaction in terminal care settings.** *Soc Sci Med* 1991, **32(1)**:83-87.
9. Kidder D: **The effects of hospice coverage on Medicare expenditures.** *Health Serv Res* 1992, **27(2)**:195-217.
10. Mor V, Kidder D: **Cost savings in hospice: final results of the National Hospice Study.** *Health Serv Res* 1985, **20(4)**:407-422.
11. Temel JS, Greer JA, Muzikansky A, Gallagher ER, Admane S, Jackson VA, Dahlin CM, Blinderman CD, Jacobsen J, Pirl WF, *et al*: **Early palliative care for patients with metastatic non-small-cell lung cancer.** *N Engl J Med* 2010, **363(8)**:733-742.
12. Christakis NA: **Timing of referral of terminally ill patients to an outpatient hospice.** *J Gen Intern Med* 1994, **9(6)**:314-320.
13. Christakis NA, Escarce JJ: **Survival of Medicare patients after enrolment in hospice programs.** *The New England Journal of Medicine* 1996, **335(3)**:172-178.
14. [http://www.ssa.gov/OP_Home/ssact/title18/1861.htm].
15. Dawes RM, Faust D, Meehl PE: **Clinical versus actuarial judgment.** *Science* 1989, **243(4899)**:1668-1674.
16. Hastie R, Dawes RM: **Rational choice in an uncertain world.** London: Sage Publications Inc; 2001.
17. The Support Investigators: **A Controlled Trial to Improve Care for Seriously Ill Hospitalized Patients: The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT).** *JAMA* 1995, **247(20)**:1591-1598.
18. Cook DJ, Guyatt GH, Jaeschke R, Reeve J, Spanier A, King D, Molloy DW, Willan A, Streiner DL, Canadian Critical Care, Trials G, *et al*: **Determinants in Canadian Health Care Workers of the Decision to Withdraw Life Support From the Critically Ill.** *JAMA* 1995, **273(9)**:703-708.
19. Hamel MB, Goldman L, Teno J, Lynn J, Davis RB, Harrell FE, Connors AF, Califf R, Kussin P, Bellamy P, *et al*: **Identification of Comatose Patients at High Risk for Death or Severe Disability.** *JAMA* 1995, **273(23)**:1842-1848.
20. Teno J, Lynn J, Connors A, Wenger N, Phillips RS, Alzola C, Murphy DP, Desbiens N, WA K: **The illusion of end-of-life resource savings with advance directives. SUPPORT Investigators. Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment.** *J Am Geriatr Soc* 1997, **45(4)**:519-520.
21. Teno JM, Lynn J, Phillips RS, Murphy D, Youngner SJ, Bellamy P, Connors AFJ, Desbiens NA, Fulkerson W, Knaus WA: **Do formal advance directives affect resuscitation decisions and the use of resources for seriously ill patients? SUPPORT Investigators. Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments.** *J Clin Ethics* 1994, **5(1)**:23-30.
22. Knaus WA, Harrell FE, Lynn J, Goldman L, Phillips RS, Connors AF, Dawson NV, Fulkerson WJ, Califf RM, Desbiens N, *et al*: **The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults.** *Annals of Internal Medicine* 1995, **122(3)**:191-203.
23. Lynn J, Knaus WA: **Background for SUPPORT.** *J Clin Epidimiol* 1990, **43(Suppl)**:1S-4S.
24. Tsalatsanis A, Hozo I, Vickers A, Djulbegovic B: **A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making.** *BMC Med Inform Decis Mak* 2010, **10**:51.
25. Vickers A, Elkin E: **Decision curve analysis: a novel method for evaluating prediction models.** *Med Dec Making* 2006, **26(6)**:565-574.

26. Djulbegovic B, Hozo I: When Should Potentially False Research Findings Be Considered Acceptable? *PLoS Med* 2007, 4(2):e26.
27. Djulbegovic B, Hozo I, Schwartz A, McMasters KM: Acceptable regret in medical decision making. *Med Hypotheses* 1999, 53(3):253-259.
28. Hozo I, Djulbegovic B: When is diagnostic testing inappropriate or irrational? Acceptable regret approach. *Med Dec Making* 2008, 28(4):540-553.
29. Hozo I, Djulbegovic B: Will insistence on practicing medicine according to expected utility theory lead to an increase in diagnostic testing? *Med Dec Making* 2009, 29:320-322.
30. Bell DE: Regret in Decision Making under Uncertainty. *Operations Research* 1982, 30:961-981.
31. Loomes G, Sugden R: Regret theory: an alternative theory of rational choice. *Economic J* 1982, 92:805-824.
32. Zeelenberg M, Pieters R: A theory of regret regulation 1.1. *J Consumer Psychol* 2007, 17:29-35.
33. Olsen LA, Aisner D, McGinnis MJ: The learning healthcare system: Workshop summary (IOM roundtable on evidence based medicine). Washington: The National Academic Press; 2008.
34. Eden J, Wheatley B, McNeil B, Sox H: Knowing what works in health care: A roadmap for the Nation. Washington: The National Academies Press; 2008.
35. McCaffery M, Beebe A: Pain: Clinical manual for nursing practice. Baltimore: W Mosby Company; 1993.
36. Gigerenzer G, Todd PM, ABC-Research-Group: Simple heuristics that make us smart. New York: Oxford University Press; 1999.
37. Schwartz LM, Woloshin S, Welch HG: Using a Drug Facts Box to Communicate Drug Benefits and Harms. *Annals of Internal Medicine* 2009, 150(8):516-527.
38. Kahneman D: Maps of bounded rationality: psychology for behavioral economics. *American Economic Review* 2003, 93:1449-1475.
39. Krantz DH, Kunreuther HC: Goals and plans in decision making. *Judgement and decision making* 2007, 52(3):137-168.
40. Rawls J: A theory of justice. Revised edition. Cambridge: Harvard University Press; 1999.
41. Astrow AB, Popp B: The Palliative Care Information Act in real life. *N Engl J Med* 2011, 364(20):1885-1887.
42. Mack JW, Weeks JC, Wright AA, Block SD, Prigerson HG: End-of-life discussions, goal attainment, and distress at the end of life: predictors and outcomes of receipt of care consistent with preferences. *J Clin Oncol* 2010, 28(7):1203-1208.

Pre-publication history
The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1472-6947/11/77/prepub

doi:10.1186/1472-6947-11-77
Cite this article as: Tsalatsanis *et al.*: Extensions to Regret-based Decision Curve Analysis: An application to hospice referral for terminal patients. *BMC Medical Informatics and Decision Making* 2011 11:77.

33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Rough Set Theory based Prognostication of Life Expectancy for Terminally Ill Patients

Eleazar Gil-Herrera, Ali Yalcin, Athanasios Tsalatsanis, Laura E. Barnes and Benjamin Djulbegovic

Abstract—**We present a novel knowledge discovery methodology that relies on Rough Set Theory to predict the life expectancy of terminally ill patients in an effort to improve the hospice referral process. Life expectancy prognostication is particularly valuable for terminally ill patients since it enables them and their families to initiate end-of-life discussions and choose the most desired management strategy for the remainder of their lives. We utilize retrospective data from 9105 patients to demonstrate the design and implementation details of a series of classifiers developed to identify potential hospice candidates. Preliminary results confirm the efficacy of the proposed methodology. We envision our work as a part of a comprehensive decision support system designed to assist terminally ill patients in making end-of-life care decisions.**

I. Introduction

ACCORDING to Medicare regulations, a patient should be referred to hospice if his/her life expectancy is less than 6 months [1]. However, despite the well-documented advantages of hospice services, terminally ill patients do not reap the maximum benefits of hospice care with the majority of them being referred to hospice either prematurely or too late. In general, premature hospice referral is translated to patients losing the opportunity to receive potentially effective treatment, which may have prolonged their lives. Conversely, late hospice referral reduces the quality of life for patients and their families. It is apparent that accurate prognostication of life expectancy is of vital importance for all parties involved in the hospice referral process (e.g. patients, their families, and their physicians).

Here, we propose a novel knowledge discovery methodology developed to identify terminally ill patients with life expectancy less than 6 months. The core of the proposed methodology is Rough Set Theory [2]. The rest of this paper describes implementation details, reports results, and discusses limitations and future directions of our work.

II. Methodology

A. Literature Review

Approaches for developing prognostic models for estimating survival for seriously ill patients range from the use of traditional statistical and probabilistic techniques [3]-[6], to models based on artificial intelligence techniques such as neural networks, decision trees and rough set methods [7]-[11]. A recent systematic review of prognostic tools for estimating survival in palliative care highlighted the lack of accurate end-of-life prognostic models [13].

Both statistics based techniques and *AI* based models rely on data that are precisely well defined. However, medical information, which represents patients records that include symptoms and clinical signs, is not always well defined and, therefore, the data are represented with vagueness [14]. Particularly, for this kind of information, it becomes very difficult to classify borderline cases in which very small differences in the value of a variable of interest may completely change categorization and therefore the following decisions can changes dramatically [15]. Moreover, the dataset is presented with inconsistencies in the sense that it is possible to have more than one patient with the same description but showing different outcomes.

In this work we propose the use of Rough Set Theory (RST) [2] to deal with vagueness and inconsistency in the representation of the dataset. RST provides a mathematical tool for representing and reasoning about vagueness and inconsistency. Its fundamentals are based on the construction of similarity relations between dataset objects from which approximate yet useful solutions are provided. In RST, the knowledge extracted from the data set is represented in the form of "if-then" decision rules where an explanation of how the final decision was derived can be traced. Clinical credibility in prognosis models depends on the ease with which practitioners and patients can understand and interpret the results [16]. Therefore, the if-then decision rule representation offers a significant advantage over "black box" modeling approaches such as neural networks.

RST has been used in a number of applications dealing with modeling medical prognosis [9]–[12]. For example, Tsumoto et al. [11], provides a framework to model medical diagnosis rules showing theoretically that the characteristics of medical reasoning reflect the concepts of approximation established in Rough Set Theory. Komorowski et al. [12], show that RST is useful to extract medical diagnosis rules to identify a group of patients for whom performing a test that is costly or invasive is redundant or superfluous in the prognosis of a particular medical condition.

In this paper we describe a RST based knowledge discovery methodology to provide a classifier that properly discriminates patients into two groups, those who survive at least 180 days after evaluation for hospice referral and those who do not. ROSETTA [17] software is used to perform the analysis described in the remainder of the paper.

Manuscript received March 26th, 2011. This work was supported in part by the Department of Army under grant #W81 XWH-09-2-0175.

E. Gil-Herrera and A. Yalcin are with the Department of Industrial and Management System Engineering, University of South Florida, Tampa, FL 33620, USA(e-mail: eleazar@mail.usf.edu, ayalcin@eng.usf.edu).

A. Tsalatsanis, L. E. Barnes and B. Djulbegovic are with the Center for Evidence Based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL 33612, USA (e-mail: atsalats@health.usf.edu, lbarnes@health.usf.edu, bdjulbeg@health.usf.edu).

B. Dataset

The dataset used in this study consists of the 9105 cases from the SUPPORT (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments) prognostic model dataset [18]. We consider all variables used in the SUPPORT prognostic model [4] as condition attributes, i.e. the physiologic variables along with the diagnosis groups, age, number of days in the hospital before entering the study, presence of cancer, and neurologic function. Attributes' names and descriptions are listed in Table I.

As the decision attribute, we define a binary variable (Yes/No) "deceases_in_6months" using the following two attributes from the SUPPORT dataset:

TABLE I
CONDITION ATTRIBUTES

Name	Description
meanbp	Mean arterial blood pressure Day 3
wblc	White blood cell count Day 3
hrt	Heart rate Day 3
resp	Respiratory rate Day 3
temp	Temperature (Celsius)
alb	Serum Albumin
bili	Bilirubin
crea	Serum Creatinine
sod	Sodium
pafi	Pa02 / (.01 * FiO2)
ca	Presence of cancer
age	Patient's age
hday	Days in hospital at study admit
dzgroup	Diagnosis group
scoma	SUPPORT coma score based on Glasgow coma scale

- *"death"* which represents the event of death at any time up to NDI date (National Death Index date: Dec 31, 1994).
- *"D.time":* number of days of follow up

The values of the decision attribute are calculated converting the *"D.time"* value in months and comparing against the attribute *"death"* as follows:

- If "D.time" < 6 months and "death" is equal to 1 (the patient died within 6 months) then "deceases_in_6months" is equal to "Yes"
- If "D.time" > 6 months and "death" is equal to 1 (the patient died after 6 months) then "deceases_in_6months" is equal to "No"
- If "D.time" > 6 months and "death" is equal to 0 (the patient did not died after 6 months) then "deceases_in_6months" is equal to "No"

C. Rough Set Theory

Based on RST, we can formally define the prognostication problem as:

$$T = (U, A \cup \{d\}) \quad (1)$$

where T represents the dataset in the form of a table. Each row represents an object and each column represents an attribute. U is a non-empty finite set of objects and the set A represents a non-empty finite set of attributes called the condition attributes. In our case, an object designates a terminally ill patient and an attribute designates each of the fifteen condition attributes that describe a patient (Table I). Also, for every attribute $a \in A$, the function a: $U \rightarrow V_a$ makes a correspondence between an object in U to an attribute value V_a which is called the value set of a.

The set T incorporates an additional attribute $\{d\}$ called the decision attribute. The system represented by this scheme is called a *decision system.*

D. Rough Set Theory Based Knowledge Discovery Process

RST based knowledge discovery process requires sequential and parallel use of various mathematical, statistical and soft computing methodologies with the objective of identifying meaningful relationships between condition and decision attributes.

The selection of specific methodologies for knowledge discovery is largely dependent on the considered dataset. We have taken the following steps in our approach:

1) Data preprocessing: If the selected table contains "holes" in the form of missing values or empty cell entries; the table may be processed in various ways to yield a completed table in which all entries are present. The data completion process for SUPPORT dataset in [18] is adopted in this work. After the preprocessing phase, the number of patients with missing information is reduced by 2 cases. Therefore, there are 9103 complete cases.

The next step in preprocessing is the discretization process. 13 out of 15 of the conditional attributes are continuous; therefore we transformed them into categorical variables. The discretization process is based on the searching of cuts that determine intervals. This process enables the classifier in obtaining a higher quality of classification rules. We found that using cut-off defined by medical experts is the best alternative for the discretization process. We consider the APACHE III Scoring System [5] for determining the cut-off for the physiologic variables along with the age variable. The remaining variables, not defined in [5] are discretized using Boolean Reasoning Algorithm [19] implemented in the ROSETTA software.

Finally, the dataset is divided randomly into training and testing sets containing 500 and 8603 cases, respectively. The training set is used in the discretization process to obtain the cut-off for the numerical attributes.

2) Reduct Generation: This step reduces the dimensionality of the dataset with the intention of removing redundant information and consequently decreases the complexity of the mining process. A reduct is the minimal set of attributes that enable the same classification as the complete set of attributes without loss of information. There are many algorithms for computing reducts for which the effect to the classification performance is critical. Since the computational complexity of the reduct generation problem is NP-hard [19], various suboptimal techniques have been proposed. In this work the dynamic reduct approach ([20-21]) is used for reduct generation.

2.1) Dynamic Reducts

Dynamic reducts algorithm aims at obtaining the most

stable sets of reducts for a given dataset by sampling within this dataset. Random samples of the testing set are selected iteratively and reducts for the samples are computed using genetic algorithms [22-23]. The reducts that most frequently appear in the samples are the most stable.

Based on the principle of the dynamic reducts technique, we have randomly selected 100 subdivisions of the training set to use for reduct generation. The actual number of patient profiles included in each subdivision of the training set varies between 50% and 90% of the training dataset. Using this approach, 229 reducts were obtained from which the set of decision rules are generated.

2.2) Using the decision attribute as condition attribute

Typically only the condition attributes are used to generate reducts. As an alternative, we included the decision attribute d in the set of condition attributes and calculated the reducts based on this scheme.

The decision attribute (deceases_in_6_months) used as a condition attribute is intended to represent the physician's estimate of life expectancy expressed in terms of the decision classes defined for this problem. Survival prognosis models that incorporate physician estimates are shown to improve both predictive accuracy and the ability to identify patients with high probabilities of survival or death [4]. In this case, 549 reducts were obtained. The next step is the induction of decision rules.

3) Rule Induction. The ultimate goal of the RST based knowledge discovery methodology is to generate decision rules, which will be used in classifying each patient as surviving or not surviving within the defined period of time. A decision rule has the form: *if A then B* ($A \rightarrow B$), where A is called the condition and B the decision of the rule. Decision rules can be thought of as a formal language for drawing conclusions from data.

The decision rules were generated based on the two aforementioned sets of reducts. After the process of reducts generation, the decision table is presented in a compact shape from which the decision rules are generated

4) Classification. Based on the set of rules generated, we can classify patients as surviving or not surviving the six-month period. However, not all rules are conclusive. Patients with profiles identical to the conditions of the rules are not decisively classified. In addition, there are situations of contradictory rules, e.g. one or more rules classify a patient as surviving and some other rules classify the same patient as dying. To overcome these problems a *standard voting* algorithm [19] is used which allows all rules to participate in the decision process and classify a patient based on majority voting.

III. Results

This section compares the performance of the classification processes where, the patients in the training dataset are classified as *survive, not survive* or *undefined* based on the induced rules and the classification process described. The results are presented in a confusion matrix form.

The accuracy of each classification model is reported in terms of Area under the Receiver Operating Characteristic curve (AUC). The best possible classification is achieved when AUC is equal to 1, while no classification ability exists when AUC is equal to 0.5.

Table 2 presents the confusion matrix for the classification model based on reducts generated on only the original condition attributes (without including the decision attribute). Table 3 shows the confusion matrix for the alternative case where the decision attribute is included in the set of condition attributes.

TABLE 2
CONFUSION MATRIX. THE REDUCTS ARE BASED ON SET A. THE CLASSIFIER PRESENTS AUC EQUAL TO 0.55 INDICATING WEAK DISCRIMINATION ABILITY.

		Predicted		
		Not survive	Survive	Undefined
Actual	Not survive	1395	1953	677
	Survive	1410	2542	626

Sensitivity = 0.64
Specificity = 0.42
AUC = 0.55

TABLE 3
CONFUSION MATRIX. THE REDUCTS ARE BASED ON SET $A = A \cup \{d\}$. THE CLASSIFIER PRESENTS AUC EQUAL TO 0.90 INDICATING GOOD DISCRIMINATION ABILITY.

		Predicted		
		Not survive	Survive	Undefined
Actual	Not survive	1999	471	1555
	Survive	312	3245	1021

Sensitivity = 0.91
Specificity = 0.81
AUC = 0.90

The dynamic reducts approach without using the decision attribute as a condition attribute shows a weak discrimination ability. However, it demonstrates a fairly high level of coverage, being able to classify around 85% of the test cases. As shown in Table 3, the classification performance in terms of AUC when using the decision attribute as a part of the condition attributes is approximately 0.90. Both the specificity and sensitivity scores are tremendously improved. However, the classification coverage in this case is reduced to 70%.

The described classification process was repeated 10 times using randomly selected samples from the dataset (again 500 cases for training and the remainder 8603 cases for testing). The overall classification performance is obtaining by averaging the AUC from each iteration. Using the original set of attributes, the overall AUC is 0.56 (SD = 0.01). Following the same, we obtained an AUC of 0.85 (SD = 0.065) for the case where the decision attribute is used as a condition attribute.

IV. Conclusions and Future Work

The SUPPORT model is the "gold standard" model for prognostication of terminally ill patients. The AUC for prediction of survival for 180 days in the SUPPORT study is 0.79, and 0.82 when SUPPORT model is combined with physician's estimates [4].

This initial exercise in applying knowledge discovery methodologies based on rough set theory shows promise in developing a reliable methodology to predict life expectancy. The baseline model using dynamic reducts presents several opprotunities for improvement:

1. Due to the limitations of the ROSETTA software, the size of the training set was limited to 500. The size of the training set may be a limiting factor to obtaining better classification accuracy and coverage considering the high number of categories associated with each attribute.
2. One area that needs to be explored is the appropriate weighting of the condition attributes in terms of their impact on the decision variable. The baseline case assumes that all physiological attributes are weighed equally. We believe that a careful weighting of the attributes by consulting an expert will greatly improve the classification accuracy of the approach.

Including the physician's estimate in the prognostication process is an important component of our future work. The classifier which uses the decision attribute as a condition attribute is intended to incorporate the professional opinion of the physician. This classifier performed much better than the baseline model and its accuracy exceeded that of the SUPPORT model. However we note that, in this approach only 70% of the test cases could be classified and more research is required to minimize the number of *undefined* cases. Furthermore, our model used the decision attribute from a retrospective study for which the decision was known with 100% accuracy. Ideally this approach should be tested on a prospective dataset and its performance compared to other soft models based on AI techniques which are a part of our future work.

Finally, it is important to remember that regardless of the accuracy of any classifier, medical decisions must take into account the individual patient preferences towards alternative forms of treatments[24]. Therefore, our intent is to incorporate our methodology into a patient-centric decision support system to facilitate the hospice referral process.

References

[1] L. R. Aiken, "Dying, Death, and Bereavement," *Allyn and Bacon*, 1985, p. 214.

[2] Z. Pawlak, "Rough Sets: Theoretical Aspects of Reasoning about Data," *Kluwer Academic Publishers*, Norwell, MA. 1992.

[3] D. W. Hosmer Jr., S. Lemeshow, "Applied Survival Analysis: Regression Modeling of Time to Event Data," *John Wiley & Sons*, Chichester, 1999.

[4] W. A. Knaus, F. E. Harrell Jr, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors Jr, et al, "The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults," Ann Intern Med. 1995, pp. 191-203. s

[5] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P.G. Bastos, C.A Sirio, D.J Murphy, T. Lotring, A. Damiano, "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults," *Chest*, vol. 100, no. 6, 1991, pp. 1619-1636.

[6] J. R. Bech, S. G. Pauker, J. E. Gottlieb, K. Klein, J. P. Kassirer, "A convenient approximation of life expectancy (The "D.E.A.LE")," Use in medical decision-making, *Am J Med.* 1982, pp. 889-97.

[7] K. J. Cios, J. Kacprzyc, "Medical Data Mining and Knowledge Discovery," *Studies in Fuzziness and Soft Computing 60,* Physica Verlag, Heidelberg, 2001.

[8] J. F. Lucas-Peter, A. Abu-Hanna, "Prognostic methods in medicine," *Artificial Intelligence in Medicine,* vol. 15, no. 2, Feb. 1999, pp. 105-119.

[9] J. Bazan, A. Osmolski, A. Skowron, D. Slezak, M. Sacauka and J. Wroblewski. "Rough Set Approach to the survival Analysis," *Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing series* , 2002, pp. 522-529.

[10] J. P. Grzymala- Busse, J. W. Grzymala-Busse, Z. S. Hippe, "Prediction of melanoma using rule induction based on rough sets," In: *Proc of SCI'01*, 2001, vol. 7, pp. 523-527.

[11] S. Tsumoto, "Modelling Medical Diagnostic Rules Based on Rough Sets," in *Proceedings of the First International Conference on Rough Sets and Current Trends in Computing (RSCTC '98),* Lech Polkowski and Andrzej Skowron (Eds.). Springer-Verlag, London, UK, 1998, pp. 475-482.

[12] J. Komorowski and A. Øhrn, "Modeling prognostic power of cardiac tests using rough sets," *Artificial intelligence in medicine*, Feb. 1999, vol. 15, no. 2, pp. 167-191.

[13] F. Lau, D. Cloutier-Fisher, C. Kuziemsky, et al. "A systematic review of prognostic tools for estimating survival time in palliative care," *Journal of Palliative Care*, 2007, vol. 23, no. 2, pp. 93-112.

[14] T. Williamson, "Vagueness," London, Routledge, 1994.

[15] B. Djulbegovic, "Medical diagnosis and philosophy of vagueness – uncertainty due to borderline cases," *Ann Intern Med.* 2008.

[16] A. Hart and J. Wyatt, "Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks," *Medical informatics,* 1990 vol. 15, no. 3, pp. 229-236.

[17] A. Øhrn, J. Komorowski, "ROSETTA: A Rough Set Toolkit for Analysis of Data," *Proc. Third International Joint Conference on Information Sciences, Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97),* Durham, NC, USA, 1997, March 1-5, vol. 3, pp. 403-407.

[18] Support Datasets Archived At ICPSR (http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets)

[19] J. G. Bazan, H. S. Nguyen, P. Synak, J. Wroblewski, "Rough set algorithms in classification problem," In: L. Polkowski, S. Tsumoto, T.Y Lin, (Eds.), "Rough set methods and applications: new developments in knowledge discovery in information systems. Studies in Fuzziness and Soft Computing," *Physica-Verlag*, Heidelberg, Germany, 2000, pp. 49-88.

[20] J. Bazan, A. Skowron, P. Synak, "Dynamic reducts as a tool for extracting laws from decision tables," *Proceedings of the Eighth International Symposium on Methodologies for Intelligent Systems. Lecture Notes in Artificial Intelligence 869*, Berlin, Springer-Verlag, 1994, pp. 346-355.

[21] J. Bazan, "Dynamic Reducts and Statistical inference," In *Sixth International conference, Information Procesing and Management of Uncertainty in Knowledge-Based Systems*, Granada, Spain, Universidad de Granda, 1996.

[22] J. Wroblewski, "Finding minimal reducts using genetic algorithms," In *Proc. Second International Joint Conference on Information Sciences*, 1995, pp. 186–189.

[23] D. E. Goldberg, "GA in search, optimization, and machine learning," *Addison-Wesley,* 1989.

[24] A.Tsalatsanis, I. Hozo, A. Vickers, B. Djulbegovic, "A regret theory approach to decision curve analysis: A novel method for eliciting decision maker's preferences and decision making," *BMC Medical Informatics and Decision making*, 2010, vol. 10, issue 51.

Tsalatsanis et al. BMC Medical Informatics and Decision Making 2010, 10:51
http://www.biomedcentral.com/1472-6947/10/51

BMC
Medical Informatics & Decision Making

RESEARCH ARTICLE Open Access

A regret theory approach to decision curve analysis: A novel method for eliciting decision makers' preferences and decision-making

Athanasios Tsalatsanis[1], Iztok Hozo[2], Andrew Vickers[3], Benjamin Djulbegovic[1,4*]

Abstract

Background: Decision curve analysis (DCA) has been proposed as an alternative method for evaluation of diagnostic tests, prediction models, and molecular markers. However, DCA is based on expected utility theory, which has been routinely violated by decision makers. Decision-making is governed by intuition (system 1), and analytical, deliberative process (system 2), thus, rational decision-making should reflect both formal principles of rationality and intuition about good decisions. We use the cognitive emotion of regret to serve as a link between systems 1 and 2 and to reformulate DCA.

Methods: First, we analysed a classic decision tree describing three decision alternatives: treat, do not treat, and treat or no treat based on a predictive model. We then computed the expected regret for each of these alternatives as the difference between the utility of the action taken and the utility of the action that, in retrospect, should have been taken. For any pair of strategies, we measure the difference in net expected regret. Finally, we employ the concept of acceptable regret to identify the circumstances under which a potentially wrong strategy is tolerable to a decision-maker.

Results: We developed a novel dual visual analog scale to describe the relationship between regret associated with "omissions" (e.g. failure to treat) vs. "commissions" (e.g. treating unnecessary) and decision maker's preferences as expressed in terms of threshold probability. We then proved that the Net Expected Regret Difference, first presented in this paper, is equivalent to net benefits as described in the original DCA. Based on the concept of acceptable regret we identified the circumstances under which a decision maker tolerates a potentially wrong decision and expressed it in terms of probability of disease.

Conclusions: We present a novel method for eliciting decision maker's preferences and an alternative derivation of DCA based on regret theory. Our approach may be intuitively more appealing to a decision-maker, particularly in those clinical situations when the best management option is the one associated with the least amount of regret (e.g. diagnosis and treatment of advanced cancer, etc).

Background

Decision making is often governed by uncertainty that inevitably affects the overall decision process. In their efforts to model uncertainty, decision theorists have proposed many methodologies with the majority of them having been based on statistics and probability [1-4], information theory and entropy [5], or possibilistic approaches such as fuzzy logic [6,7].

In clinical medical research, much effort has been invested in developing decision support systems for diagnosis and treatment of various clinical conditions such as management of infectious diseases in an intensive care unit, chronic prostatitis, or liver surgery [8-12] to name a few examples. Most of these systems are based on probabilistic prediction models. Even though prediction models have been shown to be generally superior and potentially complementary to physicians' prognostications [13-15], historically they have not fulfilled decision makers expectations to help improve decision-making. One reason for this is that most

* Correspondence: bdjulbeg@health.usf.edu
[1]Center for Evidence-based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL, USA
Full list of author information is available at the end of the article

probabilistic medical decision support systems are based on expected utility theory that humans often violate [14,16,17]. In addition, most models in medicine do not incorporate decision-makers' preferences, which in addition to having reliable evidence, is the key to rational decision-making [18-20].

The goal of this paper is to develop a novel decision-making approach that incorporates the decision maker's attitudes towards multiple treatment strategies. Our goal is addressed through the following three specific aims. First, we deviate from the traditional expected utility theory in an attempt to satisfy both formal criteria of rationality and human intuition about good decisions [18-22]. We employ regret theory, since regret is a cognitive emotion that combines both rationality and intuition, which are key elements for decision-making [22,23], to develop a novel methodology for eliciting decision makers' personal preferences. Consequently we reformulate decision curve analysis (DCA) [24,25] from the regret theory point of view to evaluate alternative treatment strategies and to integrate both evidence on prognosis and treatment with the decision maker's attitudes and preferences [26-28]. Finally, we identify circumstances under which a decision maker tolerates a wrong decision.

To implement our approach, we first compute the threshold probability at which the decision maker is indifferent between alternative actions, based on the level of regret one might feel when he/she makes a wrong decision. We then employ the regret based DCA to identify the optimal strategy for a particular decision maker. The optimal strategy is the one that brings the least regret in the case that it is, in retrospect, wrong. We also show how to employ a prediction model to estimate the probability of disease for a patient and contrast it with the decision maker's threshold probability. Finally, we incorporate the concept of acceptable regret in the decision process to identify the conditions under which the decision maker tolerates a potentially wrong decision.

Methods

Decision analysis based on regret theory

Figure 1 depicts a typical decision tree describing administration of treatment guided by a prediction model. There are two competing strategies (treat, and do not treat), and four possible outcomes as described by the combinations: treat/do not treat and necessary/unnecessary.

In Figure 1, $p = P(D+)$is the probability associated with the presence of the disease as estimated by a prediction model;$1 - p = P(D-)$is the probability associated with the absence of the disease, and, U_i, $i \in [1,4]$, are the utilities corresponding to each outcome. For example, U_1 is the utility of administering treatment to a patient who has the disease (e.g. treat when necessary), and U_2 is the utility of administering treatment to a patient who does not have the disease (e.g. administering unnecessary treatment). Note that we use the term "treatment" in the generic sense of health care intervention, which may indicate therapy, procedure, or a diagnostic test.

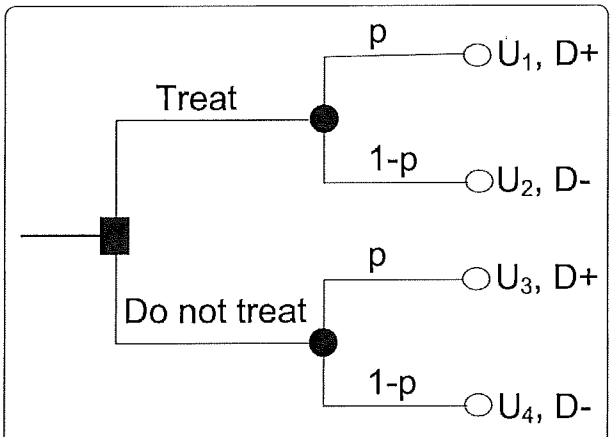


Figure 1 Decision tree for administration of treatment. In this figure, $p = P(D+)$ is the probability associated with the presence of a disease; $1 - p = P(D-)$ is the probability associated with the absence of the disease; U_i, $i \in [1,4]$, are the utilities corresponding to each outcome. Note that we use the term "treatment" in the generic sense of health care intervention, which may indicate therapy, procedure, or a diagnostic test.

The probabilistic nature of prognostication models complicates significantly the decision process. For example, if a prediction model estimates the probability of a patient having a disease equal to 40%, it is unclear whether this patient should receive treatment or not. A solution from the point of view of the classical decision theory is to employ the concept of threshold probability P_t , which is defined as the probability at which the decision maker is indifferent between two strategies (e.g. administer treatment or not) [27,29,30]. Based on the threshold concept, the patient should be treated if $p \geq P_t$ and should not be treated otherwise.

However, since in most cases decisions are made under uncertainty and can never be 100% accurate [23,26,28,31-34]. Thus, after a decision has been made one may discover that another alternative would have been preferable. This knowledge may bring a sense of loss or regret to the decision maker [23,26,28,31-34]. Regret can be particularly strong when the consequences of wrong decisions are life threatening or seriously influence the quality of the patient's life.

Formally, regret can be expressed as the difference between the utility of the outcome of the action taken and the utility of the outcome of the action that, in retrospect, should have been taken [23,26,28,31-34]. Regret

can be felt by any party involved in the decision-making process (e.g. patients receiving treatment, patient's proxies or physicians administering treatment). For the rest of this paper we assume that the decision maker is the treating physician.

We first employ regret theory to estimate the threshold probability, P_t , at which the physician is indifferent between alternative management strategies (e.g. administer treatment or not). In order to accomplish this, we describe regret in terms of the errors of (1) not treating the patient who has the disease, and (2) treating the patient who does not have the disease.

Figure 2 describes the derivation of regret associated with each strategy based on the utilities of each action's outcome. As can be noted, the regret associated with the error of not treating the patient when he/she should have received treatment (the probability of disease is $p \geq P_t$), $Rg(Rx-, D+)$, is equal to the loss in benefits of treatment. This can be expressed as the difference between the utility of receiving treatment and having the disease, and the utility of not receiving treatment and having the disease (U1-U3).

Similarly, the regret associated with treating the patient who should not have received treatment (the probability of disease is $p < P_t$), $Rg(Rx+, D-)$, is equal to harms incurred due to treatment. This can be expressed as the difference between the utilities of not having the disease and not receiving treatment, and not having the disease and receiving treatment (U4-U2). We expect no regret in the cases of correct treat/no treat decisions, $Rg(Rx+, D+) = Rg(Rx-, D-) = 0$. The difference (U1-U3) represents the consequences of not administering treatment where indicated, while (U4-U2) represents the consequences of administering treatment to a patient who does not need it. Under these assumptions, the threshold probability, P_t is equal to [27,29,30]:

$$P_t = \frac{1}{1+\frac{U_1-U_3}{U_4-U_2}} \quad (1)$$

Equation 1 effectively captures the preferences of the decision maker towards administering or not administering treatment. At the individual level, equation 1 shows how the threshold probability relates to the way the decision maker weighs false negative (i.e. failing to provide necessary treatment) vs. false positive (i.e. administering unnecessary treatment) results [24,25].

Note that the fraction $\frac{U_1-U_3}{U_4-U_2}$ is undefined for $U_4 - U_2 = 0$, which means that in this situation there is no regret associated with administering unnecessary treatment. Under these circumstances, P_t = 100%, indicating that treatment is justified only in case of absolute certainty of disease (p = 100%), a realistically unachievable goal [26].

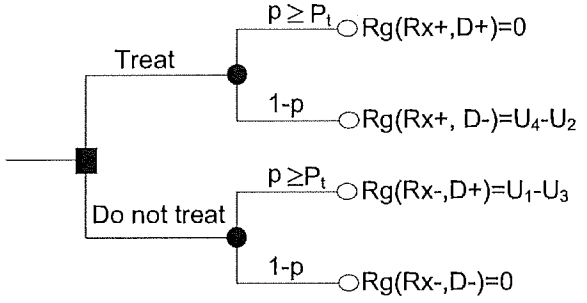


Figure 2 Regret model of the decision tree for administration of treatment. In this figure, *p*: probability of having the disease; 1-*p*: probability of not having the disease; P_t: threshold probability for treatment; *Rg*: regret associated with wrong decisions; *Rx-*: no treatment; *Rx+*: treatment; *D+*: disease is present; *D-*: disease is absent. For example, *Rg(Rx+, D-)*: regret associated with the error of treating the patient who did not have the disease.

Elicitation of threshold probability

There are numerous techniques for eliciting the decision maker's preferences regarding treatment administration [35]. None of them has been proven to be better than the other. We argue that any attempt to measure people's preferences and risk attitudes should be derived from an underlying theory of decision-making that can be applied to a problem or a class of the problems at hand. We approach elicitation of preferences by capturing people attitudes (e.g. physicians') through threshold probabilities. Normatively, a threshold probability reflects indifference between two alternative management strategies.

There are few commonly used methods to assess the value of this indifference for a decision maker such as the standard gamble, and the time trade-off [35-37]. The problem is that both standard gamble and time trade-off are time-consuming, cognitively more complex and are shown that can lead to biased estimates of people's preferences [36,37]. An alternative method is to use rating scales, such as visual analog scales (VAS), which are considerably easier to administer and better understood by the participants. The problem with analog scales, however, is that they cannot capture health state trade-offs [36,37].

The proposed method retains the simplicity of VAS but it takes into account the consequences of possible mistakes in decision-making by utilizing two visual analog scales. The first scale aims to assess the regret associated with potential error of failing to administer beneficial treatment ("regret of omission"). The second

scale measures the regret of administration of unnecessary treatment ("regret of commission"). Using these two scales we can capture trade-offs and compute the threshold probability at which a decision maker is indifferent between two alternative management strategies.

We employed the two visual analog scales with typical 100 points [35-37]anchored by no regret and maximal regret. This is modeled after pain assessment limiting the maximum possible pain that a person can experience [38]. Accordingly, we can elicit threshold probabilities by asking the physician to weigh the regret associated with wrong decisions (e.g. giving unnecessary treatment vs. failure to administer necessary treatment) using a numerical (0 to 100) scale. The questions may be narrowly defined related to specific outcomes (e.g., survival/mortality, heart attack etc.). We should, however, note that most treatments are associated with multiple dimensions, some good and some bad. This is a fundamental reason why no universally accepted method for assessment of decision-makers' preferences has been developed so far. It is very difficult, if not impossible, to accurately determine the trade-offs across multiple outcomes that can be permuted in a number of ways. A solution to this problem is to capture the decision-maker's global or "holistic" perception toward treatment. By asking questions about trade-offs in this way, we directly address both cognitive mechanisms-intuitive and deliberative- of the decision process. This, in turn, can lead to more accurate assessment of the decision makers' preferences.

For example, to elicit the physician's threshold probability, we may ask the following questions:

> *1. On a scale 0 to 100, where 0 indicates no regret and 100 indicates the maximum regret you could feel, how would you rate the level of your regret if you failed to provide necessary treatment to your patient (i.e. did not give treatment that, in retrospect, you should have given)?* [Note that the answer to this question corresponds to the (U1-U3) expression in equation 1)].
>
> *2. On a scale 0 to 100, where 0 indicates no regret and 100 indicates the maximum regret you could feel, how would you rate the level of your regret if you had administered unnecessary treatment to your patient (i.e. administered treatment that, in retrospect, should have not been given)?* [Note the answer to this question corresponds to the (U4-U2) expression in equation 1).]

For example, suppose that the physician answers 60 and 30 to the questions 1 and 2, respectively. This means that the physician considers 60/30 = 2 times worse to fail to administer treatment that should have been given than to continue unnecessary treatment. Then, the threshold probability for this physician is:

$$P_t = \frac{1}{1+\frac{U_1-U_3}{U_4-U_2}} = \frac{1}{3} = 33\%.$$

Thus, the physician would be unsure as to whether to treat or not the patient if the patient's probability of disease as computed by the prediction model was 33%. Thus, the recommended action, which is based on elicitation of the decision-maker preferences, is directly derived from the underlying theoretical model.

Regret based decision curve analysis (DCA)

Decision-makers may be presented with many alternative strategies that can be difficult to model. A simple, yet powerful approach that is based on experience of a typical practicing physician is to compare the strategy based on modeling with those scenarios when all or no patient is treated. That is, the clinical alternatives to the prediction model strategy is to assume that all patients have the disease and thus treat them all, or to assume that no patient has the disease and thus treat none [25]. In this case the clinical dilemma a physician faces when considering treatment is threefold: (1) treat all the patients ("treat all"), (2)treat no patients ("treat none"), and (3) use a prediction model and treat a patient if $p \geq P_t$ ("model").

The optimal decision depends on the preferences of the decision maker as captured by the threshold probability. We use Decision Curve Analysis (DCA) [24,25] to identify the range of threshold probabilities at which each strategy ("treat all", "treat none", and "model") is of value. Traditional DCA uses the (net expected) benefits associated with each strategy to recommend the best strategy [24,25]. In this work, we consider that the optimal strategy is the one that brings the least regret in case it is proven wrong, retrospectively.

One view about decision curves is that they should not be used in clinical practice: the researcher determines whether the decision curve justifies the use of the model in practice and then makes a simple recommendation yes or no as to whether clinicians should base their decisions on the model [39]. Another approach, which we propose here, is that threshold probabilities obtained in clinical practice should be compared against the decision curve to determine which strategy should be used (e.g. use a model, biopsy all men, biopsy no-one). This might be necessary if there is no strategy with the highest net benefit across the entire range of reasonable threshold probabilities.

Figure 3, depicts the generalized decision tree describing all of the alternative strategies. By solving the decision tree, we can estimate the expected regret associated with each strategy [23,26,28,31-34]. For example,

$$ERg[Model] = p(FN)(U_1 - U_3) + (1-p)(FP)(U_4 - U_2) \quad (2)$$

Here, *FN* (probability of false negatives) represents the conditional probability $P(p < P_t|D+)$of not treating the patient who has the disease.

FP (probability of false positives) is the conditional probability $P(p \geq P_t|D-)$of treating the patient who does not have the disease.

Similarly,

$TP = 1 - FN = P(p \geq P_t|D+)$(probability of true positives): Probability of treating the patient who has the disease.

$TN = 1 - FP = P(p < P_t|D-)$ (probability of true negatives): Probability of not treating the patient who does not have the disease.

After re-scaling the utilities by dividing each utility with the expression $U_1 - U_3$, and replacing $\frac{U_4-U_2}{U_1-U_3} = \frac{P_t}{1-P_t}$, we get the expression:

$$\begin{aligned} ERg[Model] &= p(FN) + (1-p)(FP)\frac{P_t}{1-P_t} \\ &= p(1-TP) + (1-p)(FP)\frac{P_t}{1-P_t} \\ &= P(p < P_t \cap D+) + P(p \geq P_t \cap D-)\frac{P_t}{1-P_t} \end{aligned} \quad (3)$$

For the strategies of administering treatment and not administering treatment, the expected regret is derived as:

$$ERg[Treat\ all] = (1-p)(U_4 - U_2) = (1-p)\frac{P_t}{1-P_t} \quad (4)$$

$$ERg[Treat\ none] = p(U_1 - U_3) = p \quad (5)$$

Subtracting each of these expected regrets from the expected regret of the "Treat none" (baseline) strategy we obtain the "**Net Expected Regret Difference (NERD)**":

$$\begin{aligned} NERD\ [Treat\ none,\ Model] &= ERg[Treat\ none] - ERg[Model] = \\ &= p - p(1-TP) - (1-p)(FP)\frac{P_t}{1-P_t} \\ &= p(TP) - (1-p)(FP)\frac{P_t}{1-P_t} \end{aligned} \quad (6)$$

$$\begin{aligned} NERD\ [Treat\ none,\ Treat\ all] &= ERg[Treat\ none] - ERg[\ Treat\ all] \\ &= p - (1-p)\frac{P_t}{1-P_t} \end{aligned} \quad (7)$$

$$NERD[Treat\ none, Treat\ none] = 0 \quad (8)$$

Note that these are **exactly** the same formulas as those derived by Vickers and Elkin [25] who employ the expected-utility model in "decision curve analysis" (DCA). The regret based derivation, however, is mathematically more parsimonious. The original DCA formulation required several mathematical manipulations making the simplicity of regret approach more attractive. In addition, as argued throughout the manuscript, the regret formulation may have additional decision-theoretical advantages as it enables experiencing consequences of decisions both at the emotional (system 1) and cognitive (system 2) level [23,40].

In addition to equations 6-8, we are interested in the NERD between the strategies "Treat all" and "Model":

$$\begin{aligned} NERD[Treat\ all,\ Model] &= ERg[Treat\ all] - ERg[Model] \\ &= (1-p)(TN)\frac{P_t}{1-P_t} - p(FN) \end{aligned} \quad (9)$$

The NERD equations associated with each strategy, 6-8, can be further reformulated as follows [23,25,26,28,31-34,41]:

$$\begin{aligned} NERD &= P(p \geq P_t \cap D+) - P(p \geq P_t \cap D-)\frac{P_t}{1-P_t} \\ &= \frac{\#TP}{n} - \frac{\#FP}{n} \cdot \frac{P_t}{1-P_t} \end{aligned} \quad (10)$$

Similarly, equation 9 can be re-written as:

$$\begin{aligned} NERD &= (1-p)(TN)\frac{P_t}{1-P_t} - p(FN) \\ &= P(p < P_t \cap D-)\frac{P_t}{1-P_t} - P(p < P_t \cap D+) \\ &= \frac{\#TN}{n} \cdot \frac{P_t}{1-P_t} - \frac{\#FN}{n} \end{aligned} \quad (11)$$

Equations 10 and 11 above are useful when calculating *NERD* as a function of P_t. The probabilities $P(p \geq P_t \cap D+)$, $P(p \geq P_t \cap D-)$, $P(p \geq P_t \cap D+)$, and $P(p \geq P_t \cap D-)$ are estimated as follows:

- $P(p \geq P_t \cap D+) \approx$ the number of patients who have the disease and for whom the prognostic probability is greater than or equal to P_t(with #TP = number of patients with true positive results, $P(p \geq P_t \cap D+) \approx \frac{\#TP}{n}$, where n is the total number of patients in the study).
- $P(p \geq P_t \cap D-) \approx$ the number of patients who do not have the disease and for whom the prognostic probability of disease is greater than or equal to P_t

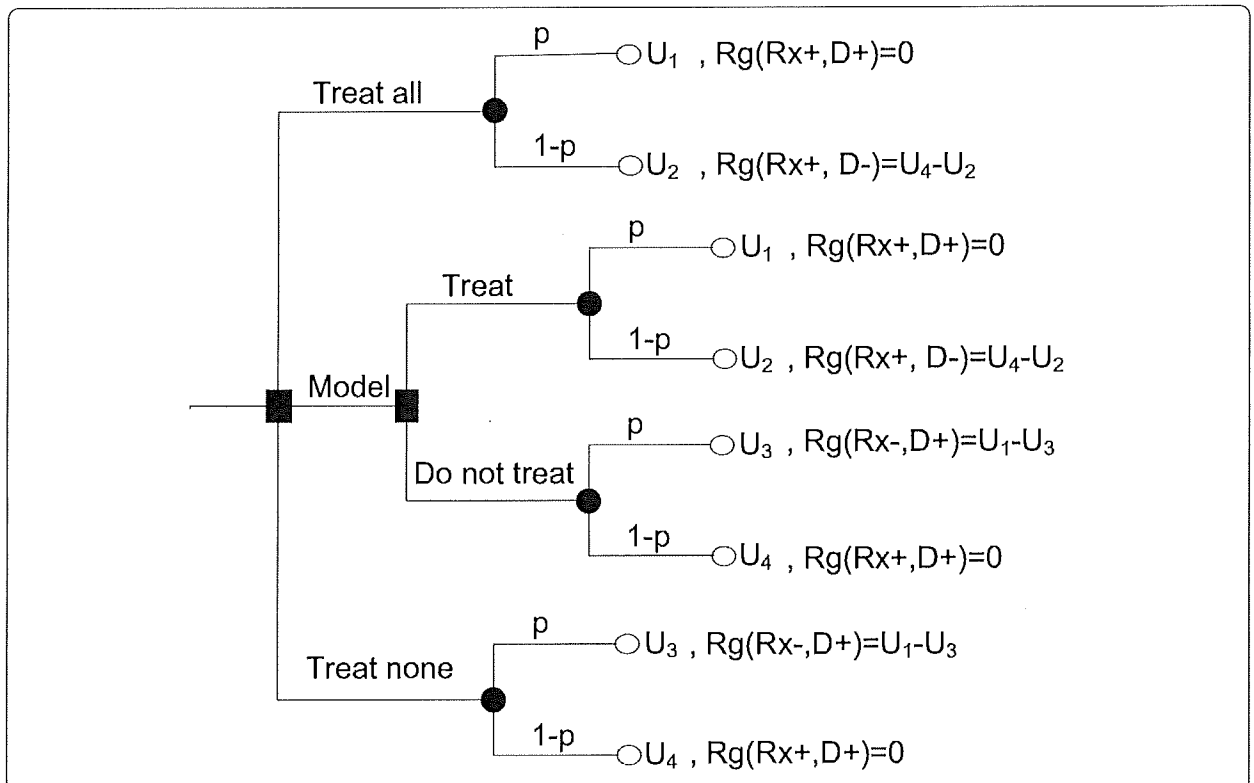


Figure 3 Generalized decision tree for administration of treatment. In this figure, *p*: probability of having the disease; 1-*p*: probability of not having the disease; *Rg*: regret associated with wrong decisions; *Rx*-: no treatment; *Rx*+: treatment; *D*+: disease is present; *D*-: disease is absent.

(with #FP = number of patients with false positive results, $P(p \geq P_t \cap D-) \approx \frac{\#FP}{n}$).

• $P(p < P_t \cap D+) \approx$ the number of patients who have the disease and for whom the prognostic probability of disease is less than P_t (with #TN = number of patients with true negative results, $P(p < P_t \cap D+) \approx \frac{\#TN}{n}$).

• $P(p < P_t \cap D-) \approx$ the number of patients who do not have the disease and for whom the prognostic probability of disease is less than P_t (with #FN=number of patients with false negative results, $P(p < P_t \cap D-) \approx \frac{\#FN}{n}$).

When computing *NERD*[*Treat none, treat all*] we assume that all patients have the disease, thus #TP is the number of people who actually have the disease and #FP is the number of people who do not have the disease but are given treatment. On the other hand, when computing *NERD*[*Treat none, Model*]from equation 10 and, *NERD*[*Treat all, Model*]from equation 11, #TP, #FP, #TN, and #FN are computed for each threshold probability assuming that a patient has the disease if the prognostic probability is greater than or equal to the threshold probability and does not have the disease, otherwise.

NERDs of each of the strategies described are plotted against different values of threshold probability. The NERD values provide information relative to **decrease in regret** when two strategies are compared against each other for a given threshold probability. If NERD = 0, this means that there is no difference in the regret between two strategies:

$$\begin{aligned} &NERD[strategy\ 1,\ strategy\ 2] = 0 \Leftrightarrow \\ &ERg(strategy1) - ERg(strategy2) = 0 \Leftrightarrow \\ &ERg(strategy1) = ERg(strategy2) \end{aligned} \tag{12}$$

If NERD > 0, this means that the second strategy will inflict less regret than the first strategy, and hence it is preferable:

$$\begin{aligned} &NERD[strategy\ 1,\ strategy\ 2] > 0 \Leftrightarrow \\ &ERg(strategy1) > ERg(strategy2) \end{aligned} \tag{13}$$

Similarly, if NERD < 0, the first strategy represents the optimal decision among the two strategies:

$$NERD[strategy\ 1,\ strategy\ 2] < 0 \Leftrightarrow ERg(strategy1) < ERg(strategy2) \quad (14)$$

The algorithm for the Regret DCA is implemented as follows:

1. Select a value for threshold probability.
2. Assuming that patients should be treated if $p \geq P_t$ and should not be treated otherwise, compute #TP and #FP for the prediction model.
3. Calculate the *NERD*(*Treat none, Model*)using equation 10.
4. Calculate *NERD*(*Treat all, Model*)using equation 11.
5. Compute the *NERD*(*Treat none, Treat all*)using equation 10 where #TP is the number of patients having the disease and #FP is the number of patients without disease who got treatment.
6. Repeat steps 1 - 6 for a range of threshold probabilities.
7. Graph each NERD calculated in steps 3-5 against each threshold probability.

Based on the Regret DCA methodology, the optimal decision at each threshold probability is derived by comparing each pair of strategies through their corresponding NERDs according to the transitivity principle (i.e., if A > B, B > C then A > C). Thus, if *NERD*(*strategy*1, *strategy*2) >*NERD*(*strategy*2, *strategy*3) > 0 then strategy 2 is better than strategy 1, and strategy 3 is better than strategy 2. Therefore, strategy 3 is the optimal strategy.

Acceptable Regret

No decision model can guarantee that the recommended strategy will be the correct one. Therefore, we can always make a mistake and recommend treatment we should not have, or fail to recommend treatment we should have administered [42]. However, there are situations where the regret resulting from a wrong decision will be tolerable. These situations are best described under the notion of acceptable regret [26,28,31]. Formally, acceptable regret,Rg_0, is defined as the portion of utility a decision maker is willing to lose/sacrifice when he/she adheres to a decision that may prove wrong [26,28,31,32]. For example, a physician may regret administering unnecessary treatment to a patient but he/she can "still live with" the consequences of this decision if she/he judged them to be trivial or inconsequential.

We assume that there is a linear relationship between the value of acceptable regret and the benefits of receiving treatment as well as the harms of receiving unnecessary treatment. This is a reasonable assumption because acceptable regret is expected to operate within a narrow range, at the lower or the upper end, of the probability scale. We define acceptable regret in terms of benefits of treatment, Rg_b, as [43] the percentage (r_b) of benefits (U_1 - U_3)the decision maker is willing to forgo if his/her decision NOT to treat was wrong:

$$Rg_0 = Rg_b = r_b B = r_b(U_1 - U_3) \quad (15)$$

Alternatively, we define acceptable regret in terms of harms of unnecessary treatment, Rg_h, as [43] the percentage (r_h) of harms (U_4 - U_2) the decision maker is willing to incur if his/her decision of treating was wrong:

$$Rg_0 = Rg_h = r_h H = r_h(U_4 - U_2) \quad (16)$$

We use the concept of acceptable regret to further refine the conditions under which the decision maker is indifferent between two strategies. Recall that these conditions have been initially captured in terms of threshold probability, which does not incorporate the sense of tolerable losses. Thus, we proceed with the following definition: Two strategies are considered *equivalent in regret* (e.g. will bring the same regret to the decision maker if they are proven wrong, in retrospect), if the absolute value of their net expected regret difference (NERD) is less than or equal to a predetermined amount of acceptable regret Rg_0. In other words, there is no difference between choosing the strategy "treat all" or "treat none" in terms of regret if:

$$|NERD(Treat\ none, Treat\ all)| \leq Rg_0 \quad (17)$$

Similarly, the strategies "model" and "treat none" are equivalent in regret if:

$$|NERD(Treat\ none, Model)| \leq Rg_0 \quad (18)$$

and the strategies "model" and "treat all":

$$|NERD(Treat\ all, Model)| \leq Rg_0 \quad (19)$$

The acceptable regret, Rg_0, can be computed using any of the two definitions described in equations 15 and 16.

We can also use equations 15 and 16 to identify the prognostic probabilities at which the decision maker would not regret the decision to which he/she is committed even if that decision may prove wrong. For instance, we are typically interested in the prognostic probability above which a physician would commit to the decision to treat a patient, and the probability below which he/she would not to treat a patient without feeling undue consequences of these decisions [28]. In other words, we are looking for the probabilities for which $ERg(Treat\ all) \leq Rg_h$, and $ERg(Treat\ none) \leq Rg_b$.

Solving the inequalities using equations 4, 5, 15, and 16 and after scaling Rg_0 by (U_1 - U_3), we obtain

$$P_{treat\ all} = 1 - r_h \tag{20}$$

Where $P_{treat\ all}$ is the prognostic probability above which the physician would tolerate giving treatment that may prove unnecessary. Similarly,

$$P_{treat\ none} = r_b \tag{21}$$

represents the prognostic probability below which the physician would comfortably withhold treatment that may prove beneficial, in retrospect.

Note that equations 20 and 21 express acceptable regret in terms of probabilities while equations 17-19 define it in terms of NERD. Hence, the outputs of these equations are not the same; rather, they complement each other.

Elicitation of acceptable regret

In most cases the decision maker does not have a complete understanding of benefits lost or harms inflicted and cannot assign a precise number to them. For this reason, we do not suggest inquiring directly about the value of r. Instead, we propose eliciting r through the decision-maker's responses to specific clinical scenarios. For example, we propose the following approach:

*Assume that you have **100 patients with the same probability of disease as the patient you are currently treating**. You need to decide whether each of these patients should receive treatment or not. Since no prediction model is 100% accurate, it is expected that you will make some mistakes in your treatment recommendations (e.g. you may recommend treatment to a patient who does not need it, or fail to recommend treatment to a patient who needs it).*

> *1. We are now interested in knowing your tolerance toward administering **unnecessary** treatment i.e. we want to learn what the magnitude of the **unavoidable error** you can live with is by inflicting potentially harmful treatment on a patient. Note that if you say that your acceptable regret is zero, this means that you can only make decision if you **absolutely certain** that your recommendation is correct.*
> *Out of the number (100-y) of patients who should have not received treatment, how many patients would you tolerate treating?* (The answer is used to compute r_h).
> *2. We are interested in knowing your tolerance toward **failing** to provide necessary treatment i.e. we want to learn what the magnitude of **unavoidable error** you can live with is by forgoing potentially beneficial treatment. Note that if you say that your acceptable regret is zero, this means that you can only make decision if you **absolutely certain** that your recommendation is correct.*
> *Out of the number (100-x) of patients who should have been treated, how many patients would you tolerate not treating?* (The answer is used to compute r_b).

It is unnecessary to ask the decision maker to answer both questions. We suggest asking only the question related to the recommendation the physician is about to make e.g. if the recommendation is about administering treatment, then the decision maker should be asked the second question, while if it is about not giving treatment, then he/she can ask the first question.

The value of acceptable regret is plotted in the regret DCA graph to visually facilitate the decision making process. At a specific threshold probability all strategies for which $|NERD| \leq Rg_0$ are considered equivalent in regret, according to the definition in the previous section.

Example

We will employ a prostate cancer biopsy example to demonstrate the applicability of our approach. Prostate cancer biopsy is an invasive and uncomfortable procedure, which can be painful and is associated with a risk of infection. However, it is often necessary for diagnosis of prostate cancer, one of the leading causes of cancer death in men.

Men are typically biopsied for prostate cancer if they have an elevated level of prostate-specific antigen (PSA). However, most men with a high PSA do not have prostate cancer. This has led to the idea that statistical models based on multiple predictors (PSA, age, family history, other markers) might be used to predict biopsy outcomes and hence aid biopsy decisions for individual patients. A physician seeing a patient with an elevated PSA has three possible options: go for biopsy, refuse biopsy or look up his probability in a statistical model and then make a decision.

We utilize an unpublished statistical model that computes probability of cancer based on the dataset described in [44] to compare each of these options. Following the algorithm described in the regret DCA section, we generate the decision curves depicted in Figure 4. This figure is used to determine the optimal strategy for different values of threshold probability. The optimization procedure is implemented in three steps where the strategies in each NERD are compared to each other as in equations 12- 14 at a specific threshold probability. For example, at threshold probability 15%:

1. *NERD(biopsy none,model)* > 0 therefore, the model is preferred to the strategy biopsy none.
2. *NERD(biopsy none, biopsy all)* > 0 therefore, the strategy biopsy all is preferred to the strategy biopsy none.
3. *NERD(biopsy all,model)* > 0 therefore, the model is preferred to the strategy biopsy none

Consequently, "model" corresponds to the optimal strategy.

Repeating the same procedure for all threshold probabilities, we can see that deciding based on the statistical model is the optimal strategy (i.e. results in the minimum expected regret) for threshold probabilities between 8% and 43%. For threshold probabilities between 42% and 95%, the optimal strategy is to biopsy no patients, while for 0% to 8% both model and biopsy all strategies are optimal.

To interpret these results, we have to consider how a typical physician values the harms of a false negative (missing a cancer) and a false positive (an unnecessary biopsy) result. If regret associated with unnecessary biopsy is felt to be worse than missing cancer, then according to equation 1, the threshold probability is greater than 50%. However, it is unlikely that a physician would consider an unnecessary biopsy to be worse than missing a cancer, so the threshold probability for biopsy must be less than 50%. Thus, a reasonable range of threshold probabilities might indeed be between 8% - 43% as suggested by our model. As the model is superior across this entire range, we can conclude that, ***irrespective of the physician's exact preferences***, making a biopsy decision based on the statistical model will lead to lower expected regret than an alternative such as biopsying all or no men. Based on discussions with clinicians, we believe that a reasonable range of threshold probability is 10% - 40%. As the regret associated with the model strategy is lowest across this entire range, we can recommend use of the model. Nonetheless, we do not have a complete sample of all physician preferences and it is possible that a physician may have a probability outside of this range.

To illustrate the applicability of the acceptable regret model, assume that the value of acceptable regret for forgoing the benefits of biopsy (equation 15) is equal to ± 0.01. Consider the case that the decision maker's threshold probability is equal to 20%. According to Figure 4, the optimal strategy should be suggested by the statistical model. However, we see that

$$\left|NERD\left(treat\ none,\ model\right)\right| < 0.01$$

which means that the strategies "biopsy none" (biopsy no patients) and "model" are equivalent in regret. Therefore, the prediction model does not offer any better information and thus, it can be disregarded.

Case Study

This section describes the overall decision process regarding prostate cancer biopsy. The process begins with elicitation of the threshold probability from the treating physician and continues with evaluation of the available strategies based on regret DCA (Figure 4). Then, if necessary, the probability of cancer based on the available prognostic model is computed and contrasted with the threshold probability. Finally, the concept of acceptable regret is employed to arrive at the strategy which is the most tolerable to the decision maker who always faces possibilities of making wrong decisions. For the remainder of this section the normal font text corresponds to the author comments. The text in **bold** and underlined font corresponds to questions to, and answers from the physician respectively. The *italic* text is notes to the reader. We demonstrate the applicability of our approach using hypothetical answers from two physicians.

The overall decision process is described as follows:

1. Interview with the physician to elicit his/her threshold probability.

a. **On the scale 0 to 100, where 0 indicates no regret and 100 indicates the maximum regret you could feel, how would you rate your level of regret if you failed to provide necessary treatment?**
Physician #1 answer: 50, Physician #2 answer: 70. *These values correspond to U_1 - U_3 from equation1.*
b. **On the scale 0 to 100, where 0 indicates no regret and 100 indicates the maximum regret you could feel, how would you rate your level of regret if you administered unnecessary treatment?**
Physician #1: 10, Physician #2: 60. *This value corresponds to U_4 - U_2 from equation 1.*

The threshold probability is equal to (equation 1): Physician #1: 16%, Physician #2: 46%.

2. Using the graph in Figure 4, identify the optimal strategy for the computed threshold probability.

Physician #1: *For threshold probability equal to 16%, the optimal decision is derived by solving the inequalities* (Figure 4, *equations 12-14):*

1. NERD(biopy all, model) > 0, *the strategy "model" is better than the strategy "biopsy all"*
2. NERD(biopsy none, model) > 0, *the strategy "model" is better than the strategy "biopsy none"*

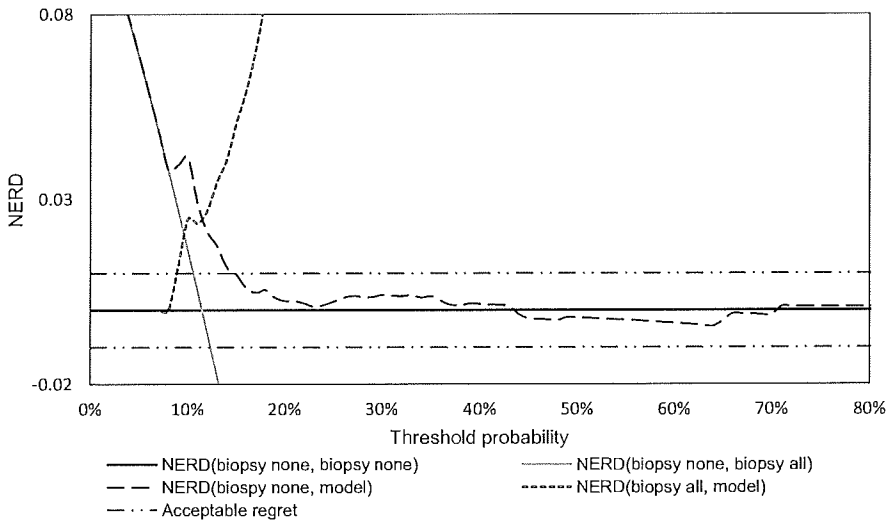


Figure 4 Regret DCA regarding biopsy to detect prostate cancer. Thin line: biopsy all patients; solid line: biopsy no patients; dashed line: prediction model. The optimal strategy is derived by the comparison of each pair of strategies from all NERDs as per equations 12-14. The statistical model is the optimal strategy for threshold probabilities between 8% and 42%. For threshold probabilities between 43% and 95%, the optimal strategy is to biopsy no patients, while for 0% to 8% both "model" and "biopsy all" strategies are optimal. The lines of acceptable regret denote the regret area in which different strategies are equivalent.For example, at threshold probability equal to 20%, the optimal strategy is acting based on the prognostic model. However, NERD(biopsy none, model) is below the acceptable regret line which indicates that the strategies "biopsy none" and "model" are equivalent in regret. Therefore the optimal strategy is to biopsy no patients as the use of model is deemed to be superfluous. Similarly,at threshold probability equal to 15%, the optimal strategy is to act based on the model and the strategies "biopsy all" and "biopsy none" are equivalent in regret. Finally, at threshold probability equal to 9%, the optimal strategies are both "model" and "biopsy all". However, since NERD(biopsy all, model) is below the acceptable regret, the strategies "biopsy all" and "model" are equivalent in regret. Therefore the optimal strategy is to biopsy all patients.

3. NERD(biopsy none, biopsy all) > 0, *the strategy "biopsy all" is better than "biopsy none".*
Therefore, the optimal strategy is the "model" which corresponds to biopsy based on the probability of cancer predicted by the statistical model. The next step is to compute the patient's probability of cancer and contrast it with the threshold probability.
Physician #2: *For threshold probability equal to 46%, the optimal decision would be the "biopsy none" strategy. In this case, even though computing the probability of cancer will not affect the physician's decision, it will help identify the circumstances under which the physician would tolerate unnecessary biopsy of the patient.*

3. Compute the cancer probability for the specific patient based on the statistical model.

a. If the cancer probability is greater than or equal to the threshold probability, then the surgeon should biopsy the patient.
b. If the cancer probability is less than the threshold probability, then the surgeon should not biopsy the patient.
Let us assume that the probability of cancer for the specific patient is equal to 20%. The threshold probability for Physician #1 is 16% (as computed in step 1). In this case, Physician #1 considers recommending biopsy. As noted in step 2b, the best strategy for Physician #2 is recommending not to biopsy any patients regardless their probability of cancer.

4. Elicitation of the level of acceptable regret.

Assume that you have 100 patients, all with probability of cancer equal to 20% (the same as your patient). This means that out of 100 patients, 20 patients will have cancer while 80 will not have cancer. You need to decide whether each of these patients should undergo biopsy or not. Since no prediction model is 100% accurate, it is expected that you will make some mistakes in your recommendations (e.g. you may recommend biopsy to a patient who does not need it, or fail to recommend biopsy to a patient who may need it).

a. The physician considers biopsy (Physician #1):
Out of the 20 patients who should be biopsied, for how many patients would you tolerate not recommending a necessary biopsy? 1.

This answer corresponds to $r_b = \frac{1}{20} = 0.05$ *and acceptable regret* $Rg_b = r_b(U_1 - U_3) = 0.05 \cdot 0.5 =$

0.025. *The optimal strategy at P_t = 16% is to use the statistical model* (Figure 4). *For P_t = 16% and Rg_h = 0.025 all NERDs are greater than acceptable regret, thus the optimal strategy remains the statistical model.*

b. The physician does not consider biopsy (Physician #2).

Out of the 80 patients who should not undergo biopsy, for how many patients would you tolerate recommending an unnecessary biopsy? 40.

The answer provided by the Physician #2 corresponds to $r_h = \frac{40}{80} = 0.50$ *and acceptable regret* $Rg_h = r_h(U_4 - U_2) = 0.5 \cdot 0.6 = 0.3$.

The optimal strategy for P_t = 46% is to biopsy no patients (Figure 4). Also, for p_t = 46% and Rg_h = 0.3, we have: $|NERD(biopsy\ none, biopsy\ all)| = |-0.639| > Rg_h$, $|NERD(biopsy\ none, model)| = |-0.003| < Rg_h$ and $|NERD(biopsy\ all, model) = 0.6364| > Rg_h$. This means that the strategies "biopsy none" and "model" are equivalent in regret. In practical terms no additional effort is justified for using the statistical model.

5. Based on equations 20 and 21, we can determine the prognostic probabilities above and under which the physician would tolerate performing an unnecessary biopsy, or not to do so when he should have done it.

a. Physician #1 considers recommending biopsy to his/her patient. Based on equation 21, the physician would tolerate not recommending a biopsy for any prognostic probability below $P_{treat\ none} = r_b$ = 5%.

b. Physician #2 considers not recommending biopsy to his/her patient. Based on equation 20, the decision maker would tolerate recommending an unnecessary biopsy for any prognostic probability above $P_{treat\ all}$ = 1 - r_h = 50%

Discussion

Currently, there is no agreed upon method for how preferences regarding multiple objectives that typically go in opposite directions (i.e. most medical interventions are associated both with benefits and harms) should be elicited. We have presented and demonstrated an approach to decision making based on regret theory and decision curve analysis. The approach presented in this paper relies on the concept of the threshold probability at which a decision maker is indifferent between strategies, to suggest the optimal decision [27,29,30]. Unlike the approaches described in the classic threshold papers [27,29,30], our approach is based on the notion that the value of threshold probability is clearly subjective and depends on the personal preferences of the decision maker. We elicit threshold probabilities based on the regret one may feel in case that the chosen strategy is proven wrong, in retrospect. Although one can narrow down the approach to specific medical outcomes, we believe that eliciting preferences in a global, holistic way is more useful if our approach is to be used in the actual practice.

We believe that the model described here has a direct practical application in overcoming many difficulties related to linking evidence with patient's preferences to arrive at the optimal decision- the issues that plagued the field of decision-making. The problem of eliciting preferences and integrating them in a coherent decision is not a simple one. We argue that the approach we are advocating here represents a contribution to the field of decision making, be should not be seen as the panacea to medical decision making. However, we anticipate our methodology to be suitable for medical decision primarily associated with trade-offs between quality and quantity of life.

Over that last couple of decades, many attempts have been made to develop the best method to take these considerations in real-life settings. Unfortunately, as explained, no approach has succeeded [35]. We believe that the reason for this is that most approaches to elicit decision maker's preferences as well as to help improve decision-making have relied on a rational framework based on expected utility theory [21]. However, modern cognitive theories (within so called dual-processing theory) have convincingly demonstrated that human decisions rely both on intuition (system 1) and analytical, deliberative process (system 2) in balancing risks and benefits in the decision-making process [22,40,45]. We believe that rational decision-making should take into account both formal principles of rationality and human intuition about good decisions [46,47]. The key is to preserve rational framework, while allowing anticipation of the effect of decision on emotions (while avoiding biases associated with intuitive thinking) [40]. One way to accomplish this is to use the cognitive emotion of regret to serve as a link between system 1 (i.e. intuitive system) and system 2 (i.e. deliberative, analytical cognitive system). By anticipating consequences of our actions and circumstances under which we can live with our mistakes, we bring together both aspects of cognition that may lead to better and more satisfactory decision-making.

Specifically, we argue that eliciting people's preferences using regret theory may be superior to using traditional utility theory because regret forces decision-makers to explicitly consider consequences of decisions. We have previously shown that we can always make errors in decision-making: recommend treatment that does not work, or fail to recommend treatment that does [26]. Therefore, we reformulated DCA from the

regret theory's point of view. Furthermore, it has been shown that the expected utility theory is often violated to minimize anticipated regret [33,34].In addition, there is substantial evidence that medical decision making aims to minimize regret associated with wrong decisions [48-50].

Moreover, while descriptive, normative, and prescriptive theories [17] tend to evaluate individual outcomes, the approach presented here evaluates all of the outcomes in a holistic manner. Our approach is consistent with Reyna's "gist" or "fuzzy trace theory" in which the decision-maker characterizes gist of each outcome to arrive at a given decision [51]. For example, consider that a decision maker is provided with a list of harms and benefits associated with each decision, as it is currently recommended by the practice guidelines panels [52]. In traditional theories, the decision maker evaluates a treatment strategy by reasoning on each of the harms and benefits associated with a given strategy. This, as discussed above, would mean integration of all multiple outcomes that often go in different directions typically within limited time-frame. Due to the complexity of these decisions, however, this approach overwhelms the decision maker as our brain capacity is limited. The regret DCA methodology quantifies the global attitudes of the decision maker towards a specific strategy without requiring separate reasoning for each of the harms and benefits. This holistic assessment occurs within the dual processing cognitive system, which evaluates collectively the harms and the benefits associated with each treatment alternative. By assessing trade-offs through both cognitive mechanisms-intuitive and deliberative-we believe that we can assess decision makers' preferences more accurately.

In general, since our method relies on the elicitation of threshold probability we recommend using our methodology for every patient. As every patient's values are different the threshold probability should indeed be patient-specific. For example, a physician may act "aggressively" for a young patient who is the father of two underage kids and less aggressively for an older patient. However, in the cancer biopsy example, it is expected that most of the patients should present with similar characteristics and therefore most physicians would settle in a small area of threshold probabilities. In this case repeating the elicitation process for every patient would be impractical. Nevertheless, this is an empirical question worthy of further investigation as alluded above.

Our approach may help reconcile formal principles of rationality and human intuitions about good decisions that may better reflect "rationality" in medical decision-making [21,32,46,47]. We hope that our theoretical work will stimulate empirical testing of the concepts outlined in this paper. Toward this end, we are currently working on developing a prescriptive computerized decision-support system to facilitate the application of the model described herein. Such a system is expected to be user friendly with built-in automatic manipulation of the complex calculations that may be off-putting to many users. We hope to report on testing of our system in the near future.

Conclusions

We have presented a decision making methodology that relies on regret theory and decision curve analysis to assist physicians in choosing between appropriate health care interventions. Our methodology utilizes the cognitive emotion of regret to determine the decision maker's preferences towards available strategies and DCA to suggest the optimal decision for the specific decision maker. We believe that our approach is suitable for those clinical situations when the best management option is the one associated with the least amount of regret (e.g. diagnosis and treatment of advanced cancer, etc).

As with any other novel theoretical work, our approach has its limitations. First, it has not been empirically tested in a clinical setting. However, we are in the process of developing the appropriate decision support tools to bring our model into clinical practice and evaluate its usefulness with actual physicians and patients. Second, the methodology presented is appropriate for single point decision making. Further investigation is required to determine the application of regret theory to decisions that re-occur over time. Finally, we assume that there is only one decision maker involved in the decision process. Nevertheless, our plan for future work includes extending our methodology to shared decision-making that will include both physician and patient in the decision process and investigate whether in practice there is a difference between preferences and choices made by physicians and their patients.

We summarize the contribution presented in this paper as follows:

1. We propose a novel method for eliciting decision makers' preferences towards treatment administration. Contrary to traditional methodologies on eliciting preferences, our method considers the consequences of potential mistakes in decisions. We propose a dual visual analog scale to capture errors of omission and errors of commission and, therefore, evaluate the trade-offs associated with each of the available strategies.
2. We have reformulated DCA from the regret theory point of view. Our approach is intuitively more appealing to a decision maker and should facilitate

decision making particularly in those clinical situations when the best management option is the one associated with the least amount of regret.

3. Finally, we utilize the concept of acceptable regret to identify the circumstances under which a decision maker tolerates a wrong decision.

We envision facilitation of the decision process in clinical settings through a computerized decision support system available at the point of care. In fact, we are in the process of developing such a system and hope to report about it soon.

Abbreviations

DCA: Decision Curve Analysis; NERD: Net Expected Regret Difference; VAS: Visual Analog Scale, p: Prognostic probability; P_t: Threshold probability; $D+/D-$: The patient has/does not have the disease; U_i: Utility corresponding to outcome i; $Rg(x)$: Regret associated with the action x; $Rx+/Rx-$: Treatment/No treatment; U_1 - U_3: Consequences of not administering treatment where indicated; U_4 - U_2: Consequences of unnecessarily administering treatment; $ERg(action)$: Expected regret associated with an action; TP,TN,FP,FN: Conditional probabilities; $\#TP,\#TN,\#FP,\#FN$: Number of TP, TN, FB, FN patients; n: Number of patients; $NERD(action1, action2)$: Net expected regret difference between actions 1 and 2; Rg_0: Acceptable regret; Rg_b: Acceptable regret as defined in terms of loses in benefits due to forgoing treatment; Rg_h: Acceptable regret as defined in terms of harms due to undergoing unnecessary treatment; r_b/r_h: Percentages of the benefits/harms a decision maker is willing to lose/incur in case of a wrong decision; $P_{treat\ all}$: The prognostic probability above which the decision maker would tolerate recommending unnecessary treatment; $P_{treat\ none}$: The prognostic probability below which the decision maker would tolerate not recommending treatment.

Acknowledgements

This work is supported by the Department of Army grant #W81 XWH 09-2-0175.

Author details

[1]Center for Evidence-based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL, USA. [2]Department of Mathematics, Indiana University Northwest, Gary, IN, USA. [3]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, NY, NY, USA. [4]H. Lee Moffitt Cancer Center& Research Institute, Tampa, FL, USA.

Authors' contributions

AT prepared the first draft, formalized the proposed methodology, and applied it into treatment administration examples; IH developed the mathematical formulation of the model; AV is the author of DCA; BD proposed the regret theory extension to DCA. All authors contributed equally in reviewing multiple versions of the paper and provided important feedback to the final version of the paper. BD is a guarantor. All authors read and approved the final draft.

Competing interests

The authors declare that they have no competing interests.

Received: 23 July 2010 Accepted: 16 September 2010
Published: 16 September 2010

References

1. Edwards W, Miles RFJ, von Winterfeldt D: **Advances in decision analysis. From foundations to applications.** New York: Cambridge University Press 2007.
2. Lindley D: **Making decisions.** New York: Willey, 2 1985.
3. Greenland S: **Probability logic and probabilistic induction.** *Epidemiology* 1998, 9:322-332.
4. Greenland S: **Bayesian Interpretation and Analysis of Research Results.** *Seminars in Hematology* 2008, 45(3):141-149.
5. Shannon C, Weaver W: **The mathematical theory of communication.** Urbana: The University of Illinois Press 1962.
6. Zimmer man H: **Fuzzy set theory and its applications.** Boston: Kluwer Academic Press, 3 1996.
7. Zimmer man H: **An application-oriented view of modelling uncertainty.** *European Journal of Operational Research* 2000, 122:190-198.
8. Schurink CAM, Lucas PJF, Hoepelman IM, Bonten MJM: **Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units.** *The Lancet Infectious Diseases* 2005, 5(5):305-312.
9. Hansen C, Zidowitz S, Hindennach M, Schenk A, Hahn H, Peitgen HO: **Interactive determination of robust safety margins for oncologic liver surgery.** *International journal of computer assisted radiology and surgery* 2009, 4(5):469-474.
10. Bratchikov OP, Korenevsii NA, Seregin SP, Dolzhenkov SD, Shumakova EA, Kotsar AG, Kriukov AA, Krivovtsev SI, Popov AV: **Automatic decision support system in prognostication, diagnosis, treatment and prophylaxis of chronic prostatitis.** *Urologiia* 2009, , 4: 44-48.
11. Bertsche T, Askoxylakis V, Habl G, Laidig F, Kaltschmidt J, Schmitt SP, Ghaderi H, Bois AZ, Milker-Zabel S, Debus J, *et al*: **Multidisciplinary pain management based on a computerized clinical decision support system in cancer pain patients.** *Pain* 2009, 147(1-3):20-28.
12. Rahilly-Tierney CR, Nash IS: **Decision-making in percutaneous coronary intervention: a survey.** *BMC Med Inform Decis Mak* 2008, 8:28.
13. Dawes RM, Faust D, Meehl PE: **Clinical versus actuarial judgment.** *Science* 1989, 243(4899):1668-1674.
14. Hastie R, Dawes RM: **Rational choice in an uncertain world.** London: Sage Publications, Inc 2001.
15. The-Support-Investigators: **A Controlled Trial to Improve Care for Seriously Ill Hospitalized Patients: The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT).** *JAMA* 1995, 247(20):1591-1598.
16. Baron J: **Thinking and deciding.** Cambridge: Cambridge University Press, 3 2000.
17. Bell DE, Raiffa H, Tversky A: **Decision making. Descriptive, normative, and prescriptive interactions.** Cambridge: Cambridge University Presspublisher 1988.
18. Djulbegovic B: **Lifting the fog of uncertainty from the practice of medicine.** *Bmj* 2004, 329(7480):1419-1420.
19. Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, Schunemann HJ: **Going from evidence to recommendations.** *Bmj* 2008, 336(7652):1049-1051.
20. O'Connor AM, Legare F, Stacey D: **Risk communication in practice: the contribution of decision aids.** *Bmj* 2003, 327(7417):736-740.
21. Djulbegovic B, Hozo I: **Health care reform & criteria for rational decisionmaking.** 2010 [http://www.smdm.org/newsletter/spring_2010/#a22].
22. Slovic P, Finucane ML, Peters E, MacGregor DG: **Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality.** *Risk Analysis* 2004, 24(2):311-321.
23. Zeelenberg M, Pieters R: **A theory of regret regulation 1.1.** *J Consumer Psychol* 2007, 17:29-35.
24. Vickers A, Cronin A, Elkin E, Gonen M: **Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers.** *BMC Medical Informatics and Decision Making* 2008, 8(1):53.
25. Vickers A, Elkin E: **Decision curve analysis: a novel method for evaluating prediction models.** *Med Dec Making* 2006, 26(6):565-574.
26. Djulbegovic B, Hozo I: **When Should Potentially False Research Findings Be Considered Acceptable?** *PLoS Med* 2007, 4(2):e26.
27. Djulbegovic B, Hozo I, Lyman GH: **Linking evidence-based medicine therapeutic summary measures to clinical decision analysis.** *MedGenMed* 2000, 2(1):E6.
28. Djulbegovic B, Hozo I, Schwartz A, McMasters KM: **Acceptable regret in medical decision making.** *Med Hypotheses* 1999, 53(3):253-259.
29. Pauker SG, Kassirer JP: **Therapeutic decision making: a cost-benefit analysis.** *N Engl J Med* 1975, 293(5):229-234.
30. Pauker SG, Kassirer JP: **The threshold approach to clinical decision making.** *N Engl J Med* 1980, 302(20):1109-1117.

31. Hozo I, Djulbegovic B: **When is diagnostic testing inappropriate or irrational? Acceptable regret approach.** *Med Dec Making* 2008, **28(4)**:540-553.
32. Hozo I, Djulbegovic B: **Will insistence on practicing medicine according to expected utility theory lead to an increase in diagnostic testing?** *Med Dec Making* 2009, 29:320-322.
33. Bell DE: **Regret in Decision Making under Uncertainty.** *Operations Research* 1982, **30**:961-981.
34. Loomes G, Sugden R: **Regret theory: an alternative theory of rational choice.** *Economic J* 1982, **92**:805-824.
35. Lichenstein S, Slovic P: **The construction of preference.** New York: Cambridge University Press 2006.
36. Stiggelbout AM, de Haes JC: **Patient preference for cancer therapy: an overview of measurement approaches.** *J Clin Oncol* 2001, **19(1)**:220-230.
37. Hunnik M, Glasziou P: **Decision-making in health and medicine. Integrating evidence and values.** Cambridge: Cambridge University Press 2001.
38. McCaffery M, Beebe A: **Pain: Clinical manual for nursing practice.** Baltimore: V.V. Mosby Company 1993.
39. Steyerberg EW, Vickers AJ: **Decision curve analysis: a discussion.** *Med Decis Making* 2008, **28(1)**:146-149.
40. Evans TSBT: **Hypothetical Thinking: Dual Processes in Reasoning and Judgement (Essays in Cognitive Psychology).** New York: Psychology Press: Taylor and Francis Group 2007.
41. Peirce CS: **The numerical measure of the success of predictions.** *Science* 1884, **4**:453-454.
42. Djulbegovic B, Frohlich A, Bennett CL: **Acting on imperfect evidence: How much regret are we ready to accept?** *J Clin Oncol* 2005, **23(28)**:6822-6825.
43. Hozo I, Schell MJ, Djulbegovic B: **Decision-Making When Data and Inferences Are Not Conclusive: Risk-Benefit and Acceptable Regret Approach.** *Seminars in Hematology* 2008, **45(3)**:150-159.
44. **Decision curve analysis.** [http://www.decisioncurveanalysis.org].
45. Kahne man D: **Maps of bounded rationality: psychology for behavioral economics.** *American Economic Review* 2003, 93:1449-1475.
46. Krantz DH, Kunreuther HC: **Goals and plans in decision making.** *Judgement and decision making* 2007, **2(3)**:137-168.
47. Rawls J: **A theory of justice. Revised edition.** Cambridge: Harvard University Press 1999.
48. Feinstein AR: **The 'chagrin factor' and qualitative decision analysis.** *Archives of internal medicine* 1985, **145(7)**:1257-1259.
49. Le Minor M, Alperovitch A, Knill-Jones RP: **Applying decision theory to medical decision-making–concept of regret and error of diagnosis.** *Methods of information in medicine* 1982, **21(1)**:3-8.
50. Hilden J, Glasziou P: **Regret graphs, diagnostic uncertainty and Youden's Index.** *Statistics in medicine* 1996, **15(10)**:969-986.
51. Reyna V: **How people make decisions that involve risk: a dual-processesapproach.** *Current Directions in Phychological Sciences* 2004, 13:60-66.
52. GRADE-working-Group: **Grading quality of evidence and strength of recommendations.** *BMJ* 2004, **328**:1490-1498.

Pre-publication history
The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1472-6947/10/51/prepub

doi:10.1186/1472-6947-10-51
Cite this article as: Tsalatsanis *et al.*: A regret theory approach to decision curve analysis: A novel method for eliciting decision makers' preferences and decision-making. *BMC Medical Informatics and Decision Making* 2010 10:51.

Title: Uncertainty about effects is a key factor influencing Institutional Review Boards' approval of clinical studies

First Author/Corresponding author
Hesborn Wao, PhD
Assistant Professor, Division of Evidence-Based Medicine and Health Outcomes Research
USF Health Morsani College of Medicine, University of South Florida
3515 East Fletcher Avenue., MDT1200 Tampa, Florida 33612
Email: hwao1@health.usf.edu | Tel: (813)974-9248

Co-Authors

Rahul Mhaskar, Ph.D.
Assistant Professor, Division of Evidence-Based Medicine and Health Outcomes Research
USF Health Morsani College of Medicine, University of South Florida
3515 East Fletcher Avenue., MDT1200 Tampa, Florida 33612

Ambuj Kumar, M.D.
Associate Professor, Division of Evidence-Based Medicine and Health Outcomes Research
USF Health Morsani College of Medicine, University of South Florida
3515 East Fletcher Avenue., MDT1200 Tampa, Florida 33612

Branko Miladinovic, Ph.D.
Associate Professor, Division of Evidence-Based Medicine and Health Outcomes Research
USF Health Morsani College of Medicine, University of South Florida
3515 East Fletcher Avenue., MDT1200 Tampa, Florida 33612

Thomas Guterbock, PhD
Professor and Director, Center for Survey Research
University of Virginia
2400 Old Ivy Road., P.O Box 40767, Charlottesville, Virginia 22904

Iztok Hozo, PhD
Professor, Department of Mathematics and Actuarial Science
Indiana University Northwest
3400 Broadway, Gary, Indiana 46408

Benjamin Djulbegovic, PhD, MD.
Distinguished Professor and Division Chief
Division of Evidence-Based Medicine and Health Outcomes Research
USF Health Morsani College of Medicine, University of South Florida
3515 East Fletcher Avenue., MDT1200 Tampa, Florida 33612

Uncertainty about effects is a key factor influencing Institutional Review Boards' approval of clinical studies

Introduction

Institutional review boards (IRBs) are locally-administered groups that undertake review of research protocols involving humans to ensure they adhere to federal regulations, adequately protect human participants' rights and welfare, and are ethically sound.[1] In the United States, the federal law mandates[2] that the Office of Human Research Protections and the Food and Drug Administration authorize IRBs to review research protocols and related materials to decide whether to approve, require modifications in planned research prior to approval, or disapprove the research. Despite the pivotal role IRBs play in research conduct, little is known about what factors influence IRBs' decision to approve or not approve a study.

If the proposed study is deemed ethical and approved by one IRB then one would expect another IRB to draw a similar conclusion. However, research show that when IRB members at multiple sites are presented with the same research proposal, their reactions vary [3, 4]. Variations have been noted in the acceptable methods for recruitment of study participants [5, 6], designation of risk level [4, 7], type of concerns expressed or changes required[5, 8-11], and more importantly, approval versus disapproval decision [11, 12]. Empirical evidence from a systematic review of 43 studies found that the same clinical study which has been approved by one IRB in the United States gets disapproved by another IRB, and vice versa [3].

Given the inconsistency in IRB's reactions to the same proposal, it is imperative to examine what factors influence IRB members' decision to approve or not approve a study protocol. According to Henderson and colleagues [13], factors such as scientific *purpose*, level of *uncertainty* about efficacy of intervention under investigation, competing *interest* of a clinician (e.g., to advance scientific knowledge vs. provide best care to patients) may influence approval decision. IRB members also may differ on approval decisions based on perceived benefit of the treatment to *current patients* versus future *patients*. According to the Belmont Report [14] and Declaration of Helsinki [15], the primary goal of research is to help develop new health intervention that will help future patients, however, research ethics require that risks to subjects posed by participation in research is justified by the anticipated benefits to the current patients [16]. Similarly, health care professionals are duty-bound not to subjugate their duties to [current] patients' best interest to the utilitarian goals for the good of others [future patients] [17, 18]. Thus, enrollment into clinical studies is justified only if it can benefit study patients more than treating them outside of the trial [17]. In sum, a number of factors may affect IRB members' approval decision. To date, there has not been any systematic investigation of factors that impact IRB members' approval decision. Accordingly, the goal of

this study was to identify factors that influence IRB members' decision to approve or not approve clinical studies.

Method

Research Design

This study comprises the qualitative component of a larger study employing mixed methods approach to examine factors that influence IRB members' decision to approve or not approve a study protocol. Specifically *partially mixed concurrent dominant status design* [19] was employed whereby quantitative and qualitative components are conducted concurrently, the quantitative component being accorded more weight in addressing the research question, and mixing occurring at the data interpretation stage.

Sample selection and participant recruitment

Potential participants included IRB members from 128 colleges and universities representing 317 IRBs which are members of the Association of American Medical Colleges and members of the Public Responsibility in Medicine and Research (PRIM&R). We obtained the list of PRIM&R members and then cross-verified these members with the list of active IRB members affiliated with 128 colleges and universities to discard duplicates. We contacted potential participants via postal mail, email and phone calls. An initial advance letter alerting recipients to the upcoming survey was used when a mailing address was available. Email was the primary mode of contact. An announcement email echoing the language of the letter was sent with the intent to reach respondents about the same time as the letter. Follow-up email reminders, along with post cards and telephone reminders, when possible, were also used. The study was approved by the University of South Florida IRB (No: 107911).

Survey development, piloting, and administration

We developed a web-based survey employing vignettes. The vignettes depicted clinical study scenarios in which uncertainties and other factors potentially influencing approval of proposed research studies were used. We used a factorial design whereby seven aspects of each scenario (Table 1) were randomly varied in 15 phrases in each vignette to produce unique vignettes for each respondent (see P1-P15 in Figure 1). At the end of each vignette participants responded to the following three questions based on a 7-point Likert-type scale (1 = Definitely Not, 2 = Most Likely Not, 3 = Likely Not, 4 = May or May Not, 5 = Likely, 6 = Most Likely, 7 = Definitely):

- extent to which IRB members believed the proposed study will generate knowledge about medical treatment that will benefit future patients
- extent to which IRB members believed the treatment in the proposed study will improve outcomes in the study patients; and
- likelihood to approve the proposed the study

Participants also responded to the open-ended question, *Please briefly describe what factors influenced your decision to approve or not approve the proposed study*. Responses to this question constitute the qualitative data upon which our report is based. Participants were asked to assume that the described study was scientifically sound and appropriate, even though not described in detail. We pilot-tested the vignettes and the web-based survey among University of South Florida IRB members and members from our study team. A sample of IRBs members at academic institutions across the United States not part of the study participants provided expert review of the vignettes and the questions asked.

We administered the survey electronically to support the intricate branching and variable randomization of the vignettes. We programmed the survey in Sensus, an Internet survey tool licensed from Sawtooth Technologies. Each respondent was presented with four pairs of vignettes, with each pair representing one of the four types of clinical studies (phase I, phase II, randomized controlled trial [RCT], and a cohort study) and one randomly selected vignette. A randomized sequence of five integers was generated, one for each respondent to be sampled, which controlled the sequence of study design types to be presented in the four pairs and the additional vignette. For each of the nine vignettes, seven random digits were used to set the values for each of the experimentally varied factors in the vignettes. Thus, a total of 68 random digits were set in advance for each respondent, so that vignette content and sequencing were fully randomized across the respondents. The random digits were edited before use to eliminate the possibility of showing to a respondent two identical vignettes within the same questionnaire. Participants were emailed the link directing them to their own copy of the survey. Participants were not directly compensated, however, we inserted a $5 gift card into all advance letters sent as a small token of appreciation to create the expectation of reciprocity.

Data analysis

We employed *thematic content analysis* whereby two members of our research team coded independently participants' significant statements in response to the open-ended question. By significant statement we refer to a statement containing a word or phrase classified under any of the nine predictors (a priori themes) or under an emergent theme. Responses were read in entirety to determine which a priori theme was implied. When none of the predictors was implied, we employed *in vivo coding* whereby a label was assigned to the response using a word or short phrase taken from that response. In vivo coding ensures that concepts stay as close as possible to respondents' own words or use their own terms to capture key elements of the construct being described. To avoid redundancy, the method of *constant comparison* was employed whereby each code was constantly compared with the preceding ones.

To glean more information, we quantitized qualitative data using theme frequency and theme intensity [20]. *Theme frequency* refers to the number of participants who cited significant

statements classified under a particular theme, expressed as a percentage of all participants. *Theme intensity* refers to the number of statements referring to a particular theme, expressed a percentage of all statements cited for all themes. Computation of these two indices is tantamount to application of quantitative analysis of qualitative data, a strategy that allows for extraction of a greater amount of information from the qualitative responses then enhancing credibility of our findings [21].

We identified two extreme cases, *approvers* and *non-approvers*. Participants who indicated that they would "definitely approve" or "most likely approve" the study were classified as approvers whereas non-approvers included participants who indicated that they would "definitely not approve" or "most likely not approve" the study. By comparing themes from these two subgroups, we aimed to develop a richer, more in-depth understanding of factors that influence approval decisions, thus lending more credibility to our findings [22]. We investigated the association between approval decision and different factors with and without taking into account the multiple observations per participant. That is, in the context on regression analysis, we adjusted for multiple observations per person by applying the random-effect logit model. Regression analyses were conducted using Stata version 13.0.[23]

Results

Participants

Our study included 42 institutions with at least one person completing the final survey. A total of 230 IRB members completed the online survey. Ninety percent (N=208) of these participants provided response to the open-ended question, *Please briefly describe what factors influenced your decision to approve or not approve the proposed study*. Table 2 presents demographic characteristics of these respondents. Majority of the respondents had experience conducting clinical research (74%) and training in research ethics (65%). In total, 2,021 significant statements were made (Cohort: 377, Phase I: 501, Phase II: 416, and RCT: 808), that is, approximately 2 statements per participant.

Factors that influence approval decision: a priori themes

Table 3 (upper panel) presents descriptions of nine a priori themes with corresponding sample significant statements obtained by coding participants' responses to the open-ended question. Uncertainty and adherence were the most frequently and intensely cited as influencing IRB members' decision to approve or not approve the proposed study (Table 4, upper panel). Uncertainty was a consistent theme across all forms of clinical study design- phase 1 through RCT and in cohort studies. Uncertainty was inferred from significant statements containing phrases such as, "uncertain/ uncertainty," "equipoise," "risk/benefit, "not sure," and/or "not known." Examples of statements using **uncertain/uncertainty** included: "...risk to participants

too high for such an uncertain benefit...," "OTC [over the counter] use of treatment W is of uncertain efficacy and safety," "Uncertainty regarding preliminary efficacy of new treatment," "Due to uncertainty indicated by experts...," "...reflects the true uncertainty required for a RCT," and "Research is done to address uncertainty. Therefore, whether the new Rx [treatment] will work in humans remains to be seen and that is exactly the purpose of research." Statements using **equipoise** included: "There is equipoise because doctors are uncertain which of the two treatments are most effective and least toxic..." and "We are nearly at equipoise. The Z arm has slightly more risk and not much chance of more benefit but we're pretty close. If 69% of docs [doctors] think Z is best, then we are obligated, as a scientific community, to see if this is true." Statements using **risk/benefit** included: "benefits outweigh risks," "... as long as study subjects know the risk/benefit ratio," "...reasonable risk/benefit," "Risk/benefit ratio is borderline acceptable," and "risk seems low and potential benefit substantial." Statements using **not sure** included: "With only a 30% rate of preventing severe disability, I am not sure whether it is worth it for patients to enroll in this study...," "A little troubled that extra blood or other safety features are not being used. Not sure this will show real change in outcomes but it may, therefore should be allowed to go ahead," "If I were a patient with RA [Rheumatoid Arthritis], I think I would enroll in such a study but I'm not sure and that makes this hard!," "QOL [quality of life] is very subjective so I would want to know more about what specifically they hope to improve and what those domains are. I am not sure if survival is an outcome here, but maybe improved QOL leads to a more positive outlook and may impact survival," and "The study drug seems somewhat benign with few adverse events - not sure about known side effects. It may prove to be beneficial to a few participants or future patients that have severe arthritis. More research is needed to test for safety and efficacy." Statements using **not known** included: "The study may involve substantial risks to subjects and the potential benefits, either for study patients or for future patients, are not known," "Optimal dose is not known. No results from human studies. Animal studies with 1% prevention rate," "Withholding a treatment p that prevents death to assess quality of life with other drug not known to prevent death," and "A 5% success rate seems very low, especially when the optimal dose is not known. Participants have to stop their current medications which seem like a lot to ask when 95% of the patients or more will probably see no improvement in their symptoms. There is also the concern that results in animals may not transfer to humans." Participants often attempted to quantify (un)certainty [24] using phrases such as "benefits outweigh risks," "High probability of improved outcome," 'study success rate," "... 100% reduction in death...," "...the risk/benefit ratio," "...reasonable risk/benefit," and "risk seems low and potential benefit substantial."

Examples of statements describing adherence included: "... patients don't have to stop standard of care treatments," "1% effect is not strong enough to warrant discontinuing meds [medication]," "...fact that other treatments are not stopped might tip the scales in favor of approval," and "... needing to stop current tx [treatment], doesn't add up to being worth it."

Factors that influence approval decision: emergent themes

Table 3 (lower panel) presents descriptions of seven emerging themes obtained by coding participants' responses to the open-ended question. Study design and harms were frequently and intensely cited to influence approval decision. Examples of comments related to study design included: "Only historic controls should be used ...;" "Design is appropriate;" "Only concern regards the choice of the cohort ...;" "The trial design appears acceptable for this study;" and "... RCT appears to be the best way to evaluate what will improve patient quality of life." The finding related to study design was surprising in that despite the fact that the participants were clearly instructed to assume that design and methods of the study were sound and appropriate; nevertheless, they still made comments suggesting that study design influenced approval decision. It should be noted that problems with study design may influence the probability of a useful outcome, changing the risk-benefit ratio. In other words, problem with study design is closely linked with the uncertainty. Examples of statements describing study harms included "minimal adverse events reported in animals given Treatment S, "100% prevention of death with minimal side effect," "Minimal adverse effects," "...relatively non-toxic," and "minimal adverse events in pts [patients] enrolled in Phase I trials reported."

Many participants indicated that additional information was needed before the approval decisions could be made. Examples of statements included: "More info needed about the gene therapy. What mechanism of action?," "There's not enough information on Treatment V ... to evaluate the study and its effects on those not responding to treatment V," "... would need more detail about the protocol to determine approval," and "Would want to understand why no experts favor treatment P and see the literature available to make a decision."

Besides study design and harms, information (i.e., the need for additional information before approval decision can be made) was frequently and intensely cited in the combined, Phase I, and RCTs subgroups as influencing the approval decision. In cohort and Phase II subgroups, however, this factor was only frequently cited as being influential.

Despite being instructed that as IRB members they have a responsibility to both current study participants and future patients, we found that approval decision was least influenced by whether a study was perceived to benefit both current and future patients (Theme #4 under Emergent themes in Table 3, lower panel). In fact, out of the 849 study vignettes that were approved, 13 (1.5%) were approved because they were perceived to benefit both current and future patients.

Identifying two extreme cases (53 approvers vs. 26 non-approvers), we found that uncertainty was frequently and intensely cited to influence approval decisions, odds ratios of 3.5 (CI: 1.3, 9.8) and 3.2 (CI: 1.1, 8.9), respectively, ignoring multiple observations per person (Table 5).

However, taking into consideration multiple observations per person, similar results were obtained for uncertainty: OR = 8.94 (CI: .93, 85.39), based on theme frequency. Furthermore, taking into consideration multiple observations per person, none of the variables was significantly associated with approval decision based on theme intensity. Overall, most factors, whether a priori themes or emergent themes, had little impact on approval decision.

Discussion

The findings from our study show that uncertainty about the intervention's effect is critical in IRB members' approval of clinical trials. This finding is consistent with prior research suggesting that clinical trials should only be conducted if there are uncertainties about the effects of competing interventions in terms of benefits and harms [18, 25, 26]. From an ethical standpoint, the development of therapeutics via testing in clinical research is justified only if uncertainties exist regarding the relative effects of new treatments versus established treatments. In the absence of such uncertainties, no research is justifiable as physicians would have prior knowledge regarding which treatment to recommend. Similarly, individual patients would not be willing to participate in such trials because it would amount to unnecessary exposure to known or unknown harms. In fact, when uncertainty is highest, randomization represents the most ethical and rational method of treatment allocation resulting in the maximum information gain. It is this "uncertainty requirement," variously referred to as uncertainty of individual physicians ("theoretical equipoise")[26], the physician and the patient ("uncertainty principle")[27, 28], the community of expert practitioners or trialists ("clinical equipoise")[29-31], or the community of patients, advocacy groups, and lay people ("community equipoise") [32, 33], that makes it necessary and possible for people to participate in RCTs by ensuring an equal chance of being randomized to whichever turns out to be the superior treatment. If the results had high likelihood of predicting treatment success (e.g., >80% that one treatment was better than the other), it would be ethically unjustifiable to deny some patients access to the superior treatment, and such a trial would ideally not be approved by IRBs. This implies that new treatment could not have been developed. From the perspective of trial development, it is actually clinical equipoise that affects design of the trial, which in turn is expected to influence the IRB's approval decision. This is the reason that our vignettes included the information on uncertainty about the proposed treatment at the level of research experts (Figure 1). Our findings are thus consistent with the notion that the IRB members actually adhere to the normative principles of equipoise.

Our results point that enrollment into clinical trials involves making choices that relate to uncertainties about hoped-for benefits and unknown harms that have not yet been observed by researchers or experienced by patients. Approval of a clinical trial represents a response of the IRB members to uncertainties, which in turn is a function of their beliefs about benefit and

harms of treatments under investigation. Our study showed that approval depends on the IRB members' beliefs about outcomes of treatments that are being tested, the beliefs that are to the large extent shaped by the uncertainties about likelihood of the patients being helped or harmed. This indicates that the IRBs should pay great attention how the information on benefits and harms are communicated to their members. We believe that improving standardization of the ways the protocol express uncertainties about treatment effects would improve the uniformity the research protocols review and the quality of the IRB review process.

An intriguing finding of our study is that IRB members seldom consider the potential benefits of a clinical study to both current and future patients in the decision to approve or not approve the study. What this finding suggests is not clear. Given that majority of IRB members in our sample had experience conducting clinical research (74%), one would expect that these IRB members would consider the potential benefits of clinical trials to both current and future patients as being important in their approval decision. Perhaps, despite IRB members' experience in clinical research, few of them consider the dual benefits of participation in clinical trials to current and future patients as being significant in trial approval decision. If that is the case then training of IRB members should emphasize on the importance of this factor in the approval decision because IRB members have a responsibility to both current study participants and future patients. Future researchers should investigate the relationship between IRB composition (e.g., in terms of experience in conducting clinical research, whether the IRB member is a scientist or non-scientist, etc.) and approval decision.

The finding that only 65% of the IRB members in our sample reported having had training in research ethics was puzzling. We suspect that participants understood "training in research ethics" to imply formal training only. Additionally, the routine training by IRB members might not be mandatory and therefore not all members might attend regularly.

Whereas information (i.e., the need for additional information before approval decision can be made) was not as influential in approval decision as adherence, research design and harms, it should be noted that information is related to uncertainty. That is, lack of knowledge or the need for additional information suggests increased uncertainty. Thus, if this factor is included under uncertainty, uncertainty emerges as the most influential factor in IRB approval decision. Our finding suggests the importance of communicating clearly information about the study design, current and expected benefit and harms and other aspects of trial useful to IRB members to consider in their decision to approve or not approve clinical studies.

An interesting finding to note is that, most of the a priori factors (e.g., potential benefit to current participants or to future patients, condition being addressed, competing interest of a clinician, procedures, scientific purpose, and therapy type) that previous research suggest may influence approval decision, were not supported by our study which is consistent with earlier

research suggesting that substantial variation exists in IRB assessment of standard protocols [4]. While it is argued that "IRB members are most effective when they are conversant with the federal rules and regulations," [34] our findings suggest that whether the study conduct follows regulatory measures or not seem to have minimal impact on approval decisions. The take-home message, especially for epidemiological researchers, is to carefully consider factors that impact IRB approval decision when preparing research protocols.

Our findings are not without limitations. Our analysis is based on nine factors identified from the literature and seven themes that emerged from data analysis. As the comments of the participants indicated, this may not be an exhaustive list of factors that may influence approval decision, which in turn may limit our conclusions. Nevertheless, the factors we took into account are overwhelming considered by the majority of the authors as the key factors that, at least, normatively must be paid attention to during IRB approval process [13]. We also cannot exclude the possibility that familiarity with the literature on IRB decision-making might have unconsciously predisposed us to a *confirmation bias* (i.e., the tendency for interpretations and conclusions based on new data to be overly congruent with prior findings) especially where rival themes were absent. However, to counteract this, not only did we determine emergent themes, to facilitate in-depth understanding of all themes, both the frequency and intensity effect sizes also were determined and compared with the quantitative findings.

In conclusion, we present the first comprehensive analysis of the most important factors affecting IRB members' research protocols approval. We found that perceived uncertainty about benefits and harms of the treatment proposed in a research protocol is a key driver of the study approvals. This, in turn, calls for improved standardization in the communications of information on benefits and harms in the research protocols considered by the IRBs.

Acknowledgements

This study was supported in part by a grant from the U.S. Department of Defense USAMRMC NO: W81XWH-09-2-0175 (PI: Djulbegovic). The authors thank IRB members at academic institutions across the United States not part of the study participants who provided expert review of the vignettes and the questions asked in the web-based survey as well as those who completed the final online survey. The authors are solely responsible for the interpretation and reporting of data in this article.

References

1. Stepke, F.L., *Standards and operational guidance for ethics review of health-related research with human participants.* Acta Bioethica, 2012. **18**(1): p. 131-131.
2. HHS, *Code of Federal Regulations, TITLE 45 PUBLIC WELFARE, Department of Health and Human Services, Part 46 PROTECTION OF HUMAN SUBJECTS.*
3. Abbott, L. and C. Grady, *A systematic review of the empirical literature evaluating IRBs: what we know and what we still need to learn.* Journal of empirical research on human research ethics : JERHRE, 2011. **6**(1): p. 3-19.
4. Mansbach, J., et al., *Variation in institutional review board responses to a standard, observational, pediatric research protocol.* Academic emergency medicine : official journal of the Society for Academic Emergency Medicine, 2007. **14**(4): p. 377-80.
5. Clark, S., et al., *Feasibility of a national fatal asthma registry: more evidence of IRB variation in evaluation of a standard protocol.* The Journal of asthma : official journal of the Association for the Care of Asthma, 2006. **43**(1): p. 19-23.
6. Silverman, H., S.C. Hull, and J. Sugarman, *Variability among institutional review boards' decisions within the context of a multicenter trial.* Critical Care Medicine, 2001. **29**(2): p. 235-241.
7. McWilliams, R., et al., *Problematic variation in local institutional review of a multicenter genetic epidemiology study.* JAMA : the journal of the American Medical Association, 2003. **290**(3): p. 360-6.
8. Greene, S.M., et al., *Impact of IRB requirements on a multicenter survey of prophylactic mastectomy outcomes.* Annals of Epidemiology, 2006. **16**(4): p. 275-278.
9. Helfand, B.T., et al., *Variation in Institutional Review Board Responses to a Standard Protocol for a Multicenter Randomized, Controlled Surgical Trial.* Journal of Urology, 2009. **181**(6): p. 2674-2679.
10. Sherwood, M.L., et al., *Unique challenges of obtaining regulatory approval for a multicenter protocol to study the genetics of RRP and suggested remedies.* Otolaryngology-Head and Neck Surgery, 2006. **135**(2): p. 189-196.
11. Stark, A.R., J.E. Tyson, and P.L. Hibberd, *Variation among institutional review boards in evaluating the design of a multicenter randomized trial.* Journal of Perinatology, 2010. **30**(3): p. 163-169.
12. Stair, T.O., et al., *Variation in institutional review board responses to a standard protocol for a multicenter clinical trial.* Academic Emergency Medicine, 2001. **8**(6): p. 636-641.
13. Henderson, G.E., et al., *Clinical trials and medical care: defining the therapeutic misconception.* PLoS Med, 2007. **4**(11): p. e324.
14. *Ethical principles and guidelines for the protection of human subjects of research [The Belmont Report]*, 1979, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Department of Health, Education, and Welfare.
15. *WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects*, 2008, World Medical Association.
16. (OHRP), O.o.H.R.P., *Institutional Review Board Guidebook, Chapter III, Basic IRB Review*.
17. Foster, C., *The ethics of medical research on humans.*2001, Cambridge: Cambridge University Press.
18. Djulbegovic, B., *Articulating and responding to uncertainties in clinical research.* The Journal of medicine and philosophy, 2007. **32**(2): p. 79-98.
19. Leech, N.L. and A.J. Onwuegbuzie, *A typology of mixed methods research designs.* Quality & Quantity, 2009. **43**(2): p. 265-275.

20. Wao, H.O., R.F. Dedrick, and J.M. Ferron, *Quantitizing text: using theme frequency and theme intensity to describe factors influencing time-to-doctorate.* Quality & Quantity, 2011. **45**(4): p. 923-934.
21. Baldwin, A.L., *Personal structure analysis: A statistical method for investigating the single personality.* Journal of Abnormal and Social Psychology, 1942. **37**: p. 163-183.
22. Patton, M.Q., *Qualitative evaluation and research methods*. 3 ed2001, Newbury Park, CA: Sage Publications.
23. *Stata Statistical Software: Release 13*, 2013, StataCorp LP: College Station, TX.
24. Djulbegovic, B., *Uncertainty and Equipoise: At Interplay Between Epistemology, Decision Making and Ethics.* American Journal of the Medical Sciences, 2011. **342**(4): p. 282-289.
25. Miller, F.G. and H. Brody, *A critique of clinical equipoise. Therapeutic misconception in the ethics of clinical trials.* The Hastings Center report, 2003. **33**(3): p. 19-28.
26. Fried, C., *Medical Experimentation: Personal Integrity and Social Policy*1974, Amsterdam, The Netherlands: North Holland Press.
27. Maurer, B.T., *The principle of uncertainty in medical practice.* JAAPA : official journal of the American Academy of Physician Assistants, 2012. **25**(6): p. 61.
28. Sonnenberg, A., *A medical uncertainty principle.* The American journal of gastroenterology, 2001. **96**(12): p. 3247-50.
29. Freedman, B., *Equipoise and the ethics of clinical research.* The New England journal of medicine, 1987. **317**(3): p. 141-5.
30. Gelfand, S., *Clinical equipoise: actual or hypothetical disagreement?* The Journal of medicine and philosophy, 2013. **38**(6): p. 590-604.
31. Weijer, C., S.H. Shapiro, and K. Cranley Glass, *For and against: clinical equipoise and not the uncertainty principle is the moral underpinning of the randomised controlled trial.* BMJ, 2000. **321**(7263): p. 756-8.
32. Gifford, F., *Community-equipoise and the ethics of randomized clinical trials.* Bioethics, 1995. **9**(2): p. 127-48.
33. Mhaskar, R., B.B. B, and B. Djulbegovic, *At what level of collective equipoise does a randomized clinical trial become ethical for the members of institutional review board/ethical committees?* Acta informatica medica : AIM : journal of the Society for Medical Informatics of Bosnia & Herzegovina : casopis Drustva za medicinsku informatiku BiH, 2013. **21**(3): p. 156-9.
34. Eissenberg, T., Panicker, S., Berenbaum, S., Epley, N., Fendrich, M., Kelso, R., Penner, L., Simmerling, M., *IRBs and Psychological Science: Ensuring a Collaborative Relationship*. 2004.

Table 1

Variable Included in Questionnaire Vignettes

Variable		Value (Description)
1	Condition	Metastatic pancreatic cancer (deadly cancer, affect survival) or rheumatoid arthritis (disabling arthritis, affect quality of life)
2	Uncertainty	Effects of the treatment/intervention, ranging from 1% to 100%
3	Therapy type	May be over the counter, gene therapy, or other
4	Study procedures	A study that requires additional testing/treatment, including invasive procedures or a study that does not require testing/treatment
5	Adherence to protocol	A study that requires stopping treatment in progress (strict protocol) or a study that does not require stopping treatment in progress
6	Scientific purpose	Either an explanatory trial (under ideal conditions of a research) or pragmatic (in a real medical practice)
7	Interest of a clinician	Either to advance scientific knowledge (scientific focus) or to provide best care for patients (patient care focus)

Table 2

Demographic Characteristics of Participants

Parameter	Number	Percent
Gender		
Female	118	57
Male	90	43
Age (years)		
30-44	26	12
45-59	63	30
60-74	45	22
75+	4	2
Not Reported	70	34
Rank		
Assistant Professor	21	10
Associate Professor	36	17
Full Professor	34	16
Instructor	5	3
Not Applicable	27	13
Other	14	7
Not Reported	70	34
Experience (years)		
1 - 5	56	27
6 - 10	51	25
11 - 15	20	10
16 - 20	7	3
20+	5	2
Not Reported	69	33

Table 3

Themes, Descriptions/Definitions, and Sample Significant Statements

Theme	Description of a Theme (Italicized) and Sample Significant Statement (In Quotes)
A priori themes	
1. Future benefit	*Knowledge gained from the study can benefit future patients* "...**benefit to future patients** seems clearer" and "There may be benefits for future patients"
2. Current benefit	Knowledge gained from the study can benefit current patients (study participants) "Given no high **benefit of current treatment**, risk to participants is too high ..."
3. Condition	*Condition being addressed (metastatic pancreatic cancer vs. rheumatoid arthritis)* "Pancreatic cancer is lethal..." and "The disease is not as fatal as pancreatic cancer"
4. Uncertainty	Level of effect of the treatment or intervention (e.g. expressed as % or risk/benefit ratio) "Due to **uncertainty** indicated by experts..."
5. Interest	*To advance scientific knowledge or to provide best care to patients* " This study is for science and not for the patient."
6. Adherence	*Whether the study adheres to protocol (e.g., require stopping concurrent treatments) or not* "...that participants **will not have to stop concurrent** Rheumatoid Arthritis treatment"
7. Procedures	*Whether the study requires additional testing/treatment and procedures or not* "... will not have **additional procedures** lessens the risks associated with the study..."
8. Purpose	*Whether the study is conducted under ideal conditions or in a real medical practice* "...if the 80% is proven correct, even with the **ideal conditions of the current study.**"
9. Therapy type	Whether over the counter, gene therapy, or other medications is used in the study " Gene therapy has significant risks and the expected future beneficiaries are few (1%)"
Emergent themes	
1. Information	Additional information is needed before approval decision can be made " Before approving the study, I would want **additional information**..."
2. Regulatory	Whether the study conduct follows regulatory measures or not " No FDA approval on treatment Z" and "Acceptable with adequate informed consent"
3. Authority	Participant defers approval decision to senior IRB members "I would consider approving the study based on the information provided at the full board meeting from members with more expertise than I have"
4. Benefit both	Knowledge gained from the study can benefit both current and future patients " If something could be learned for future patients without harming current patients,..."
5. Coercive	There is a possibility that participants are coerced into participating in the study "You cannot talk about preventing death in metastatic cancer!! That is coercive language."
6. Design	*Extent to which the study design is appropriate for the proposed research* "...**no randomization to treatment**" and "don't need control cohort in phase 1 or 2 trials."
7. Harms	*Extent to which harms to participants in the study is addressed* "The **safety parameters of this study are improved** relative to the study of Treatment T"

Table 4

Theme "Frequency" and "Intensity" (Expressed as %)

Themes	Combined		Cohort		Phase I		Phase II		RCT	
	Freq. 1542†	Int. 2102‡	Freq. 290†	Int. 377‡	Freq. 368†	Int. 501‡	Freq. 308†	Int. 416‡	Freq. 576†	Int. 808‡
A priori themes										
1) Future benefit	1.1	0.8	1.4	1.1	0.5	0.4	0.3	0.2	1.7	1.2
2) Current benefit	1.3	1.0	0.7	0.5	1.1	0.8	1.0	0.7	1.9	1.4
3) Condition	7.5	5.5	6.9	5.3	4.3	3.2	9.1	6.7	8.9	6.3
4) Uncertainty	**34**	**25**	**36**	**28**	**35**	**26**	**32**	**24**	**34**	**24**
5) Interest	5.3	3.9	6.2	4.8	4.9	3.6	4.9	3.6	5.4	3.8
6) Adherence	**24**	**17**	**20**	**15**	**27**	**20**	**26**	**19**	**22**	**16**
7) Procedures	7.1	5.2	8.3	6.4	6.5	4.8	6.5	4.8	7.3	5.2
8) Purpose	2.1	1.5	0.3	0.3	1.6	1.2	0.6	0.5	4.0	2.8
9) Therapy type	3	2.2	2.4	1.9	3.8	2.8	2.6	1.9	3.0	2.1
Emergent themes										
1) Information	**15**	**11**	**12**	9.0	**15**	**11**	**12**	8.7	**19**	**14**
2) Regulatory	2.4	1.8	2.1	1.6	3.3	2.4	3.6	2.6	1.4	1.0
3) Authority	0.2	0.1	0.3	0.3	0.3	0.2	0	0	0.2	0.1
4) Benefit both	1.9	1.4	1.7	1.3	1.4	1.0	2.3	1.7	2.1	1.5
5) Coercive	0.5	0.3	0	0	0.5	0.4	0.3	0.2	0.7	0.5
6) Design	**17**	**12**	**17**	**13**	**18**	**13**	**19**	**14**	**14**	**10**
7) Harms	**14**	**11**	**15**	**11**	**13**	9.2	**16**	**12**	**14**	**10**
8) No response	**20**	-	**24**	-	**20**	-	**21**	-	**17**	

Note: **Freq.** = Frequency; **Int.** = Intensity; **Combined** = All participants who responded to the vignettes; † = Total number of vignette responses; ‡ = Total number of themes cited; Frequency or intensity ≥ 10% is in bold face type font

Table 5

Theme Frequency and Intensity for Approvers Versus Non-Approvers

	Frequency			Intensity		
	Approver (N_R = 53)	Non-approver (N_R = 26)	OR	Approver (N_S = 53)	Non-approver (N_S = 53)	OR
	%	%	OR (95% CI)	%	%	OR (95% CI)
A priori themes						
1) Future benefit	2	0	-	2	0	-
2) Condition	11	15	0.7 (0.2, 2.7)	11	17	0.6 (0.2, 2.5)
3) Uncertainty	57	27	3.5 (1.3, 9.8)*	57	29	3.2 (1.1, 8.9)*
4) Interest	11	0	-	11	0	-
5) Adherence	25	27	0.9 (0.3, 2.6)	25	29	0.8 (0.3, 2.3)
6) Procedures	9	12	0.8 (0.2, 3.6)	9	13	0.7 (0.2, 3.3)
7) Therapy type	4	8	0.5 (0.1, 3.5)	4	8	0.4 (0.1, 1.8)
Emergent themes						
1) Information	6	19	0.3 (0.1, 1.2)	6	21	0.2 (0.2, 1.7)
2) Regulatory	0	8	-	0	8	-
3) Authority	0	4	-	0	4	-
4) Benefit both	4	0	-	4	0	-
5) Design	19	27	0.6 (0.2, 1.9)	19	29	0.6 (0.2, 1.7)
6) Harms	21	8	3.1 (0.6, 15.4)	21	8	2.9 (0.6, 14.2)

Note: * = significant at .05 level; N_R = number of respondents; N_s = number of significant statements; OR = odds ratio; CI = 95% confidence interval.

Figure 1 Sample vignette showing 15 randomly varied phrases (marked P1 to P15)

Instructions

Please review the following vignette carefully and answer the three questions following the vignette. ***In your review of this vignette, please assume that the design and methods of the study are scientifically sound and appropriate, even though they are not described in detail. Do not focus on providing criticism of information that is lacking (i.e., sample size, inclusion criteria) [P1].*** Instead, consider the overall description of the study. Please remember that as an Institutional Review Board member, you have a responsibility to both participants in the study and people who may later be given the treatment being tested if it is found to be safe and effective. We refer to people who may get the treatment at a later time as "future patients."

Metastatic pancreatic cancer [P2] is a ***deadly [P3]*** condition. Most metastatic pancreatic cancer patients live for three to four months after diagnosis. Currently, ***death [P4]*** can be prevented in half of those patients suffering from Metastatic pancreatic cancer by prescribing an existing treatment (Treatment Y). An ***alternative treatment [P5]*** (Treatment X) has been developed during the last 3 years to treat another disease. Patients who have taken Treatment X and patients who have taken Treatment Y have experienced minimal adverse events.

Some health care providers have been using Treatment X off label to treat patients with metastatic pancreatic cancer. Treatment X is not approved by the Food and Drug Administration (FDA) for the treatment of metastatic pancreatic cancer. Doctors are uncertain which of these two treatments is the most effective and the least toxic. If either drug fails to treat a patient's metastatic pancreatic cancer the patient will most certainly ***die [P6]***.

A group of investigators have submitted a proposal to your Institutional Review Board proposing a randomized controlled trial (RCT) testing whether treatment X is better than treatment Y in improving patients' ***survival [P7]***. That is, their goal is to learn which of these two treatments improves ***survival [P8]*** better ***under the ideal conditions of a research study and not necessarily [P9]*** in a real medical practice. The proposed trial is a well-designed RCT in which treatment assignment is determined by chance (random assignment). In other words, there is a ***50% [P10]*** chance of patients receiving Treatment X versus Treatment Y. The enrolled participants will be advised to follow the study treatment protocol, which will require stopping all concurrent metastatic pancreatic cancer treatment(s) upon initiation of the study. The RCT targets 40-60 year old adult patients.

The results of a survey of 100 leading experts on metastatic pancreatic cancer are available to you. The survey indicates that ***50 experts [P11]*** favor treatment X and ***50 experts [P12]*** favor treatment Y. Participation in this trial ***does [P13]*** requires additional testing and procedures over and beyond routine care. This ***might [P14]*** include frequent blood tests, x-rays, and possibly endoscopy or other invasive procedures. The investigators want to enroll patients in this study because their primary goal is to ***advance scientific knowledge [P15].***

Tsalatsanis A, Hozo I, Djulbegovic B. Empirical testing of regret-based threshold model in the end-of-life care. 37th Annual Meeting of Society of Medical Decision-Making, October 18-21, 2015 St Louis, MO.

Tsalatsanis A, Hozo I, Djulbegovic B. Empirical evaluation of regret and acceptable regret model. 36th Annual Meeting of Society of Medical Decision Making, Miami, October 19-22, 2014.

J.M. Hernandez, A. Tsalatsanis, B. Djulbegovic, V. Velanovich, "Regret theory modeling in pancreatic adenocarcinoma" (poster). Annual Cancer Symposium of the Society of Surgical Oncology Orlando, Florida, Mar 21-24, 2012.

R. Mhaskar, B. Miladinovic, A. Tsalatsanis, A. Mbah, A. Kumar, K. Sehwan, R. Schonwetter, B. Djulbegovic, "External validation of prognostic models in terminally ill patients" (poster). ASH Annual Meeting and Expo, San Diego, California, Dec 10-13, 2011. Abstract published in Blood, vol. 118 (21), 2011.

I. Hozo, A. Tsalatsanis, A. Vickers, B. Djulbegovic, "A regret theory approach to decision curve analysis" (poster). Annual Meeting of Society for Medical Decision Making (SMDM), Toronto, Canada, Oct 18-21, 2010.

B. Djulbegovic, J. Beckstead, A. Tsalatsanis, R. Mhaskar, A. Flynn, O. Fabelo, H. Tuch, A. Kumar, E. Pathak, I. Hozo, P. Jacobsen, "Anticipatory regret of commission but not omission leads to low post-decisional regret in terminally ill patients" (oral). Biennial European Meeting of SMDM.

Empirical testing of the regret-based threshold model in end of life care

Athanasios Tsalatsanis[1], Iztok Hozo[2], Benjamin Djulbegovic[1,3] for the Helping Clarify Peoples' Choices in the End-of-Life Setting Research Group

[1]University of South Florida, Tampa, FL USA
[2]Indiana University Northwest, Gary, IN, USA
[3]H. Lee Moffitt Cancer Center. Tampa, FL, USA

Purpose: The threshold model represents one of the most important advances in medical decision-making but it has never been empirically tested in real-life. We aimed to empirically test the regret-based threshold model in the end of life setting where patients choose between hospice care and treatment.

Methods: According to the regret-based threshold model there must be some probability of death (pDeath) at which patients should be indifferent (Pt) between hospice care (Hospice) and continuing treatment targeted at their disease (Rx). The model predicts that if pDeath>Pt, patients should choose hospice; if pDeath<Pt, they should opt for Rx. We tested these predictions by interviewing 134 terminally ill patients facing Rx vs. Hospice decisions. We determined Pt by eliciting regret of omission (i.e. losing benefits of hospice care) and regret of commission (i.e. incurring harms from unnecessary treatment) using a dual visual analogue scale [1]. We estimated pDeath over 6-month interval using the Palliative Performance Scale (PPS) and adjusted PPS prognostic model. We compared the regret-based threshold model recommendation to the patients' choice at two different time frames: immediately after the interview and one month after the interview to study the patients' preferences and actual choice of care. We used Cramer V (effect size) to calculate the strength of agreement between the model recommendations and the patients' preferences and actual choice, respectively.

Results: We observed statistically significant agreement between the model recommendations and the patients' stated preferences ($p<0.0001$). Out of 134 patients 111 (83%) agreed with the model recommendations immediately after the interview, 6 patients (4%) disagreed, and 17 (13%) were unsure about their preferences (figure). This converts into very large effect size (0.84). 111/134 patients were approached one month after the interview to determine what type of care the patients actually chose: 59 (53%) chose according to the model recommendations; 39 (35%) chose a different option than the model's recommendation; and 13 (12%) patients remained unsure. While the association remains statistically significant ($p=0.0067$), the effect size dropped to 0.21 indicating medium effect.

Conclusions: The regret-based threshold model strongly predicts what patients think they would want (preferences) and moderately predicts the patients' actual choice. This is the first empirical study testing the threshold model in a real-life setting.

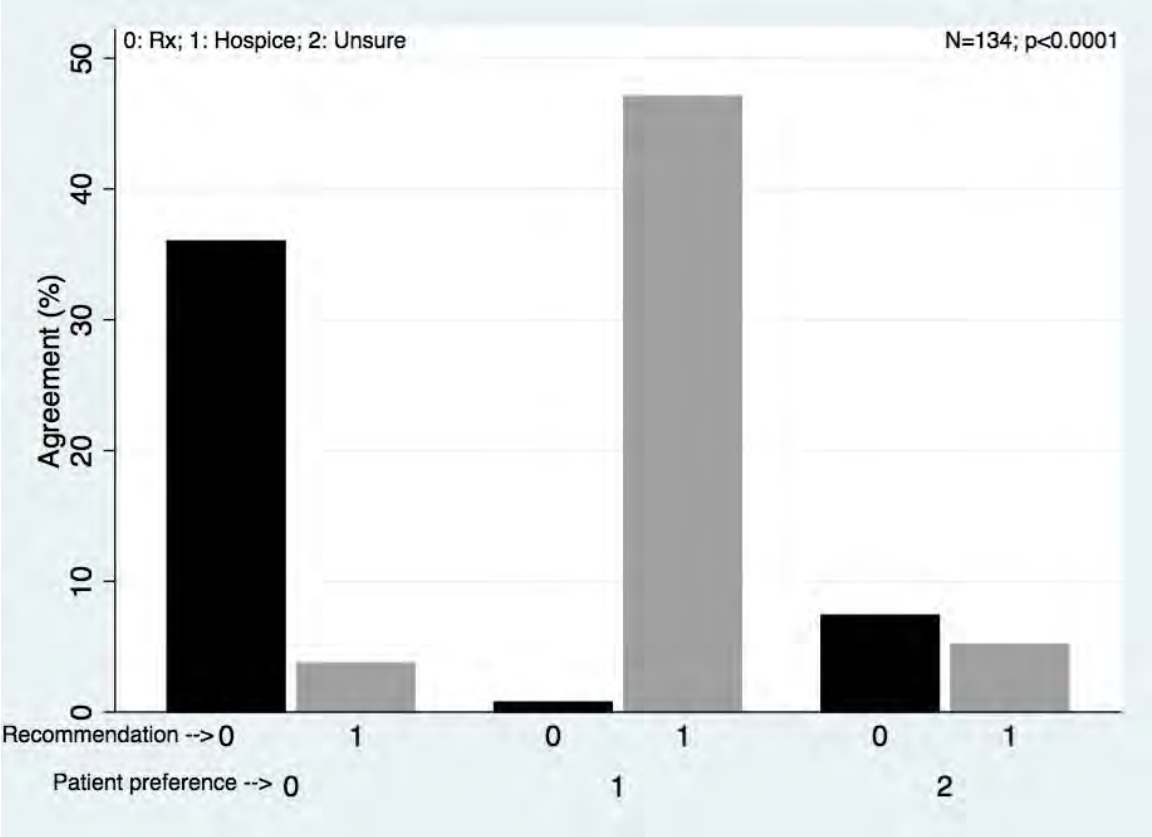


Figure. Agreement between patient preferences and regret threshold model recommendation. Effect size 0.84.

References

1. Tsalatsanis A, Hozo I, Vickers A, Djulbegovic B. A regret theory approach to decision curve analysis: A novel method for eliciting decision makers' preferences and decision-making. BMC Medical Informatics and Decision Making 2010;10:51.

Empirical evaluation of the acceptable regret model of medical decision- making

Athanasios Tsalatsanis[1], Iztok Hozo[2], Benjamin Djulbegovic[1]
(1) University of South Florida, Tampa, FL, (2) Indiana University, Gary, IN

Purpose: The acceptable regret model postulates that under specific circumstances decision makers may tolerate wrong decisions [REF]. The purpose of this work is to empirically evaluate the acceptable regret model of decision-making in end-of-life care settings, where terminally ill patients are torn between decision to seek curative treatment vs. accepting hospice/palliative care.

Methods: We conducted interviews with 23 patients enrolled in the study assessing their preferences about treatment choice in the end-of –life setting. After providing information about their life expectancy and assessing overall regret of potentially wrong choice [CITE DVAS], we elicited the patients' level of acceptable regret. We first assessed the patients' tolerance for wrongly accepting hospice care and then measured the patients' tolerance toward continuing unnecessary treatment. For the purposes of our study, a treatment was considered unnecessary if a patient dies within 6 months of the treatment. Accepting hospice care was considered a wrong decision if a patient survives longer than 6 months after the referral to hospice. We use the elicited acceptable regret levels to compute: 1) the probability of death above which a patient would tolerate wrongly accepting hospice care and 2) the probability of death below which the patient would tolerate unnecessary treatment.

Results: We found that the median probability of death above which a decision maker would tolerate wrongly accepting hospice care is 98%, while the median probability of death below which a decision maker would tolerate unnecessary treatment is 4%. The results indicate that patients do require high level of certainty to make a decision to be comfortable with potentially wrong decision (<4% and >98%, respectively). We also found that there is no statistical association between the values of acceptable regret related to wrong hospice referral (mean=1.68; SD=2.3; min=0; max=7.28) are different than the values associated with unnecessary treatment (mean=1.27; SD=1.97; min=0; max=6.58) ($p>0.05$). This finding shows that patients accept wrong competing decisions at similar levels.

Conclusions: We have elicited preliminary empirical data that corroborated the acceptable regret theory. Our results may explain why has been so difficult to provide palliative care in the end of life setting.

Findings:

1.
Probability of death below which a patient would tolerate unnecessary treatment

Variable	Obs	Mean	Std. Dev.	Min	Max
Ptreatmentall	23	.0512536	.0585191	0	.2358326

2.
Probability of death above which a patient would tolerate wrong hospice referral

Variable	Obs	Mean	Std. Dev.	Min	Max
Phospiceall	23	.9335378	.0862881	.7039439	1

3.
As shown by 1 and 2, the majority of patients will accept wrong decisions when events are almost certain (~2% or ~98%).

4.
Acceptable regret Rx

Variable	Obs	Mean	Std. Dev.	Min	Max
RgRh	23	1.274084	1.970447	0	6.584267

5.
Acceptable regret hospice

Variable	Obs	Mean	Std. Dev.	Min	Max
RgRb	23	1.685227	2.348523	0	7.2814

6.
From 4 and 5, there is no significant difference between the means or SD for RgRh and RgRb (t-test)

7.
There is no significant difference between the means or SD for RgRh and RgRb enrolled to hospice subgroups (t-test)

ABSTRACT

Background: Pancreatic adenocarcinoma is a disease which has both an overall poor prognosis and treatment which can carry significant morbidity and even mortality. Because of this, physicians are frequently left with a difficult choice of being passive and allowing the disease to run its course, or be aggressive with the potential for significant high-quality survival, but also the potential of significant short-term complications. Physicians may regret either choice if the outcome is poor. Regret theory serves as a framework linking both rationality and intuition, in order to determine an optimal course of action physicians and surgeons caring for patients with pancreatic adenocarcinoma.

Methods: Based on previous work, we generated a Cox regression model using four variables which have been shown to impact overall survival for patients with pancreatic adenocarcinoma: Pretreatment stage, resection, pretreatment vitality, and pathologic stage. We then evaluated the model using regret based decision curve analysis (regret DCA), which translates the probability estimated by the model to a decision by taking into account the decision maker's preference expressed in terms of threshold probability. By taking into account decision-maker's preferences, the analysis modeled three possible choices: always perform surgery on all patients, never perform surgery in any patient, and act according to the prediction model.

Results: 153 consecutive patients with pancreatic adenocarcinoma of all stages were ~~seen and~~ evaluated by a single surgeon at a tertiary referral center. Preoperative stage (p=0.005, CI 1.19-2.27), resection (p=0.007, CI 0.27-0.82), vitality (p<0.001, CI 0.96-0.98) and pathologic stage (p<0.001, CI 3.06-16.05) were each independent predictors of overall survival. As seen in figure 1, for a threshold probability <50% (decision maker considers failing to operate more regretful than unnecessary surgery ~~treatment~~), the least regretful decision and therefore the optimal decision is to operate on all patients regardless of the model prediction. For a threshold probability >50% (decision maker considers ~~administering~~ unnecessary surgery more regretful than failing to operate), the optimal decision is to follow the recommendations of the model and contrast the threshold probability (i.e., decision-makers' preferences) to the model's prediction of death. Specifically, for a threshold probability > 50% treatment should be administered if the probability of death as predicted by the model is greater than the threshold probability, and treatment should be held if the probability of death is less than the threshold probability.

Conclusions: Regret theory in conjunction with regret based decision curve analysis provides a novel perspective in treatment in decisions by incorporating the decision-maker's preferences with his/her

estimates about benefits and harms of performing surgery. We used this framework to analyze the decision to operate on patients with pancreatic adenocarcinoma. In this setting, surgery is always the preferable choice when regret of failing to operate is greater than the regret of unnecessary surgery. Conversely, when the regret of performing unnecessary surgery is greater, the surgeon should adhere to the survival prediction models.

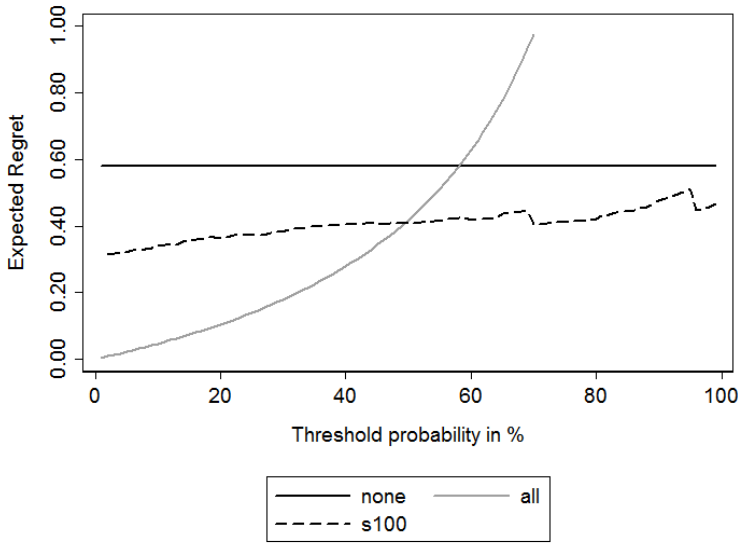


Figure 1. Regret based Decision Curve Analysis for a Cox model developed to predict survival for patients suffering from pancreatic adenocarcinoma. In regret DCA, the optimal decision is the action which will bring the least amount of regret in case it is proven wrong, in retrospect. Therefore, for threshold probabilities less than 50%, the optimal decision is to operate on all patients while for threshold probabilities greater than 50% the optimal decision corresponds to utilization of the prediction model.

External Validation of Prognostic Models in Terminally Ill Patients

Program: Oral and Poster Abstracts
Session: 901. Health Services and Outcomes Research: Poster III

Monday, December 12, 2011, 6:00 PM-8:00 PM
Hall GH (San Diego Convention Center)

Rahul Mhaskar, MPH, PhD[1*], *Branko Miladinovic, PhD*[2*], *Athanasios Tsalatsanis, PhD*[2*], *Alfred Mbah, PhD*[2*], *Ambuj Kumar, MD, MPH*[3], *Kim Sehwan, PhD*[4*], *Ronald Schonwetter, MD*[5*] *and Benjamin Djulbegovic, MD, PhD*[6]

[1]*Center for Evidence-Based Medicine, University of South Florida, Tampa, FL*
[2]*USF, Tampa, FL*
[3]*University of South Florida, College of Medicine, Center for Evidence Based Medicine, Tampa*
[4]*HPC healthcare, Tampa, FL*
[5]*HPC Healthcare, Tampa, FL*
[6]*Center for Evidence-Based Medicine & Health Outcomes Research, University of South Florida, Tampa, FL*

Background: Over one million Medicare beneficiaries receive hospice care annually. However, besides the well documented advantages of hospice, many Americans do not enjoy maximum benefit from the hospice care. The fundamental reason for this is related to the inappropriate and poorly timed referral of terminally ill patients to hospice. As a result, many patients die within a few days of referral, while some live many years after the referral was made. Improvement in the accuracy of prognosis translates into superior quality of care. Predictions based on statistical modeling have been shown to be superior to physicians' prognostication. However, very few of these statistical models have been externally validated in terminally ill patients. Here we report the external validation of 5 most commonly used prognostication models in a cohort of terminally ill patients: 1) declining exponential approximation of life expectancy (DEALE) 2) study to understand prognoses and preferences for outcomes and risks of treatments (SUPPORT), 3) adjusted palliative performance scale (PPS), 4) adjusted Karnofsky performance scale index (Karnofsky) and 5) adjusted eastern cooperative oncology group performance status (ECOG).
Methods: We retrospectively extracted data from 590 deceased patients enrolled in Tampa Bay Lifepath Hospice and Palliative Care starting January 2009 and going backwards to validate the prognostic models. Two research assistants extracted all data necessary to populate the model variables and two faculty members randomly checked 25% of the data for accuracy. The models were tested against observed survival duration. PPS, Karnofsky and ECOG risk scores were predicted using a flexible family of Royston-Parmar parametric models and adjusted for age, gender and presence of cancer. We utilized several metrics to assess the performance of these models. Specifically, we used the Brier score and scaled Brier score (which is very similar to the Pearson correlation coefficient R^2), the area under the receiver operating characteristic curve (AUROC), and the Hosmer-Lemshow goodness-of-fit p-value (HL).

Results: Brier scores were consistently below the non-informative level of 0.25 and AUROC significantly higher than the non-informative level of 0.5 for the adjusted PPS, Karnofsky and ECOG models (table 1). The HL p-value was consistently greater than 0.1 only for PPS. SUPPORT and DEALE models did not predict fit our data well for survival at day one and month one, two and six. The AUROC takes a value close to 0.5, even though the Brier scores were relatively low and HL p- value greater than 0.05, this value is significantly close to 0.5 for SUPPORT and DEALE models (table 1).

Conclusion: None of the prognostication models accurately predicated survival among our cohort of terminally ill patients. However, PPS consistently performed best in predicting survival in terminally ill patients followed by Karnofsky and ECOG.

Table 1 Discrimination and model performance statistics in survival prognostication					
	Model performance metrics	**PPS**	**Karnofsky**	**ECOG**	**DEALE**
Day 1	Brier	0.089	0.089	0.106	0.106
	Brier Scaled	8.80%	4.30%	NA	NA
	AUROC (95%CI)	0.747 (0.68, 0.813)	0.747 (0.693, 0.810)	0.709 (0.648, 0.771)	0.526 (0.467, 0.58
	HL p-value	0.26	0.17	0.11	0.44
Day 3	Brier	0.179	0.178	0.293	_
	Brier Scaled	16%	15.30%	0%	_
	AUROC	0.768 (0.726, 0.810)	0.778 (0.737, 0.818)	0.719 (0.676, 0.761)	_
	HL p-value	0.29	0.04	0.57	_
Day 6 (Median)	Brier	0.199	0.194	0.363	_
	Brier Scaled	20.10%	19.40%	NA	_
	AUROC (95%CI)	0.775 (0.739,0.816)	0.787 (0.749, 0.823)	0.721 (0.679, 0.764)	_
	HL p-value	0.43	0.008	0.44	_
Day 10	Brier	0.179	0.183	0.253	_
	Brier Scaled	26.70%	25.90%	4.70%	_
	AUROC (95%CI)	0.795 (0.757, 0.834)	0.798 (0.761, 0.836)	0.742 (0.697, 0.786)	_
	HL p-value	0.15	0.01	0.61	_
Day 30	Brier	0.122	0.127	0.095	0.099
	Brier Scaled	37.80%	37.50%	12.30%	2.44%
	AUROC (95%CI)	0.781 (0.725,0.838)	0.787 (0.734, 0.839)	0.722 (0.651, 0.794)	0.52 (0.467, 0.57
	HL p-value	0.62	0.25	0.484	0.92
Day 60	Brier	0.084	0.088	0.04	0.045
	Brier Scaled	47.40%	47.70%	18.70%	16.10%
	AUROC (95%CI)	0.745 (0.653, 0.837)	0.781 (0.689, 0.871)	0.739 (0.62, 0.858)	0.543 (0.468, 0.61
	HL p-value	0.29	0.32	0.32	0.32
Day 180	Brier	0.041	0.05	0.006	0.1
	Brier Scaled	63.20%	58.60%	31.20%	NA
	AUROC (95%CI)	0.55 (0.452, 0.648)	0.51 (0.314, 0.71)	0.59 (0.355, 0.83)	0.7 (0.386, 1)
	HL p-value	0.59	0.54	0.22	_

CI: confidence interval, H-L: Hosmer-Lemshow statistics, AUROC: area under the receiver operating characteristic curve

A regret theory approach to decision curve analysis

Iztok Hozo[1], Athanasios Tsalatsanis[2], Andrew Vickers[3], Benjamin Djulbegovic[2,4]

[1]Indiana University,[2]University of South Florida, Center for Evidence-based Medicine, [3]Memorial Sloan Kettering Cancer Center, NY and H. [4,]Lee Moffitt Cancer Center & Research Institute, Tampa, FL

Purpose: Decision curve analysis (DCA) has been proposed as an alternative method for evaluation of diagnostic tests, prediction models, and molecular markers [MDM 2006;26:565]. We re-formulated DCA from the regret theory point of view to take into consideration decision consequences [MDM 2008;28:540].

Methods: First, we constructed a classic decision tree describing three decision alternatives: treat, do not treat, and treat/no treat based on a predictive model. We then computed the expected regret for each of these alternatives as the difference between the utility of the action taken and the utility of the action that should have been taken, in retrospect. We evaluated the expected regret(s) for the tree using different weightings regarding regret associated with "omissions" (e.g. failure to treat) vs. "commissions" (e.g. treating unnecessary). For any pair of strategies, we measure the difference in net expected regret (NERD), by subtracting the expected regret of each alternative from the other. Finally, using the concept of acceptable regret-the range at which potential regret associated with wrong decisions becomes acceptable-we identified the circumstances where acting on a given strategy will be acceptable even if the decision was wrong.

Results: We first showed that NERD is equivalent to net benefits as described in the original DCA. The regret re-formulation of the original DCA model showed an asymmetry in decision-making. That is, the decision-maker seems to weigh (=the threshold probability at which a decision-maker is indifferent between two actions) regret associated with failure to treat much higher than the regret related to unnecessary treatment. This is because the decision-maker weights true positive results to a much greater extent than false-positive results. Similarly, different attitudes toward omissions vs. commissions identified different circumstances when the decision-maker can "live with" regret even if the decision was wrong, in retrospect. The symmetry in the decision-making was re-instated when the weighting for false-positive and false-negative results was identical.

Conclusions: We present an alternative derivation of the DCA based on regret theory. Under assumptions of unequal weighting, the regret approach generates identical results to the original DCA. The regret approach may also be intuitively more appealing to a decision-maker, particularly in those clinical situations when the best management option is the one associated with the least amount of regret (e.g. treatment of advanced cancer, etc).

ANTICIPATORY REGRET OF COMMISSION BUT NOT OMISSION LEADS TO LOW POSTDECISIONAL REGRET IN TERMINALLY ILL PATIENTS

Benjamin Djulbegovic1 , Jason Beckstead2 , Athanasios Tsalatsanis1 , Rahul Mhaskar1 , Alexandra Flynn3, Orlando Fabelo4 , Howard Tuch5 , Ambuj Kumar1 , Elizabeth Pathak6 , Iztok Hozo7 , Paul Jacobsen3 1 Medicine, Division of EBM, University of South Florida, Tampa, United States, 2 College of Nursing, University of South Florida, Tampa, United States, 3 Health Outcomes & Behavior, H. Lee Moffitt Cancer Center and Research Institute, Tampa, United States, 4 Clinical Research, TGH, Tampa, United States, 5 Palliative Medicine, University of South Florida, Tampa, United States, 6 Medicine, Division of EBM, University of South Florida, Tampa, United States, 7 Mathematics, Indiana University, Gary, United States Background:

A substantial body of the literature in nonmedical domains indicates that, when making their choices, the people tend to be regret averse: they anticipate regret to avoid post-decisional regret. In the terminal phase of life, the patients face regret of omission (failure to accept referral to hospice, which may help alleviate unnecessary suffering) and regret of commission (continuation of potential harmful treatment). We sought to determine which of these types of regrets represents a regret-minimizing strategy leading to lower post-decisional regret. Methods: Thirty-two patients in the terminal phase of their lives consented to the study. The preferences were elicited using the regret-based Dual Visual Analog Scale within the framework of the regret threshold model to help a patient decide between hospice referral versus continuation of disease-oriented treatment. We also recorded the patients' actual choices regarding further management and compared them with the model's recommendations. One month later the Decision Regret Scale was administered to obtain the assessment of post-decisional regret. Results: Scores on the Decision Regret Scale showed high coefficient reliability (Cronbach's alpha = 0.953). Scale scores had a moderate negative correlation with regret of commission ($r = -0.494$; $p = 0.004$) but a weaker correlation with regret of omission ($r = -0.229$; $p > .05$). Regret of omission had a negative correlation with the threshold probability ($r = -0.646$; $p < 0.001$) (probability of death at which the patient is indifferent between hospice referral and further treatment), whereas regret of commission had a positive correlation with the threshold probability ($r = 0.480$; $p = .005$). The relationships appear to be moderated by whether the model's recommendations were consistent with patient's actual choices, although this was not tested due to the restricted size of our sample. Conclusions: Our results obtained in the setting with the patients undergoing high-stakes decisions indicate that the riskier strategy of continuing potentially harmful treatment with low chance of benefits is associated with high anticipated regret of commission, which, however, in turn led to lower post-decisional regret. This result is consistent with the notion that terminally ill patients would rather accept the riskier ("leave no stone unturned") than the safer strategy to feel less regret.